UNIVERSITAT POLITÈCNICA DE VALÈNCIA

CARMA 2024

6 th International Conference on Advanced Research Methods and Analytics

June 26-28, 2024

Valencia, Spain

ADE Facultat d'Administració i Direcció d'Empreses /UPV

dec Departamento de Economía y Ciencias Sociales

Ivie

GENERALITAT VALENCIANA

LIDL

Congress UPV
6th International Conference on Advanced Research Methods and Analytics (CARMA 2024)

The contents of this publication have been evaluated by the Program Committee according to the procedure described in the preface. More information at http://www.carmaconf.org/

Scientific Editors

Josep Domenech
María Rosalía Vicente
Pablo de Pedraza

Cover design by Gaia Leandri

Publisher

2024, Editorial Universitat Politècnica de València
www.lalibreria.upv.es / Ref.: 6726_01_01_01

ISBN: 978-84-1396- 201-6

ISSN: 2951-9748

DOI: https://doi.org/10.4995/CARMA2024.2024.19017

Preface

Josep Domenech [1], María Rosalía Vicente [2], Pablo de Pedraza [1]
[1]Dept. Economics and Social Sciences, Universitat Politècnica de València, Spain. [2]Dept. Applied Economics, Universidad de Oviedo, Spain.

Abstract
This volume presents selected papers from the Sixth International Conference on Advanced Research Methods and Analytics (CARMA 2024), hosted by Universitat Politècnica de València. Highlighting the intersection of Economics and Social Sciences with Internet and Big Data, the conference showcased innovative research methodologies that reshape these fields. Key topics included public opinion mining, web scraping, digital transformation, and the governance of digital economies. Through keynote speeches, tutorials, and presentations, CARMA 2024 offered insights into the latest trends and tools, fostering a platform for academic exchange and advancing the frontiers of digital research methodologies.

Keywords: *Public Opinion Mining, Web scraping; Machine Learning Forecasting, Digital Economy, Internet Econometrics, Digital transformation, Bibliometrics.*

1. Preface to CARMA 2024

This volume proudly presents the selected papers of the Sixth International Conference on Advanced Research Methods and Analytics (CARMA 2024) hosted by the Universitat Politècnica de València, Spain from 26 to 28 June 2024. This sixth edition consolidates CARMA as a unique forum where Economics and Social Sciences research meets Internet and Big Data. CARMA offers researchers and practitioners an ideal environment to exchange ideas and explore how Internet and Big Data sources and methods can address challenges in Economics and Social Sciences and drive societal changes following digital transformation.

The selection of the scientific program was directed by Maria Rosalia Vicente and Pablo de Pedraza, who led an international team of 46 scientific committee members representing institutions worldwide. Following the call for papers, the conference received 76 paper submissions from all around the globe. All submissions were reviewed by the scientific committee members under a double-blind review process. Finally, 49 papers were accepted for oral presentation during the conference, ensuring a high-quality scientific program. It covers a wide range of research topics on the Internet and Big Data, including public opinion mining, web scraping, search engine data, natural language processing, machine learning forecasting,

and digital economy governance, among others. The program also featured nine poster presentations with promising work-in-progress research.

In this edition of CARMA, the keynote presentations offered compelling insights into emerging methodologies in social science. The program featured Mercè Crosas, who leads the Computational Social Sciences initiative at the Barcelona Supercomputing Center. Her address explored the new digital methodologies enhancing the interplay between social sciences and data science. Another highlight is the presentation by Zhijing Jin, affiliated with the Max Planck Institute and ETH Zurich, who discussed the transformative potential of natural language processing and large language models in social science research. Lastly, Jan Kinne, a postdoc at the Center for European Economic Research and Harvard University, shared his expertise on the innovative use of web data for economic research. These speakers underscore the conference's focus on the integration of advanced data techniques in the study of economic and social phenomena.

The conference also featured five tutorials designed to provide hands-on experience with cutting-edge methodologies. These tutorials covered a range of topics vital for researchers in the field. The first dealt with computational text analysis methods, offering practical applications for theory building. Another tutorial focused on topic modeling techniques, exploring various algorithms and their suitability for different data types. The session on LinkedIn data highlighted how researchers can use this platform for business and economic research. Another tutorial addressed best practices for obtaining accurate data from Google Trends, including theoretical insights and case studies. Lastly, a tutorial on bibliometric literature reviews integrated Big Data and AI to navigate complex information landscapes, equipping researchers with advanced methodologies to enhance their literature review processes.

Moreover, the conference included a session on career opportunities in Data Science and Business Intelligence, featuring a panel talk and open hour with Lidl Data & AI executives. This session was targeted at students from various academic backgrounds, providing insights into career paths, application tips, skill development, and certification guidance.

The conference organizing committee would like to thank all who made this sixth edition of CARMA a great success. Specifically, thanks are indebted to the authors, scientific committee members, invited speakers, session chairs, reviewers, presenters, sponsors, supporters, and all the attendees. We are particularly grateful to IVIE and Lidl for their generous and continuous support. Our final words of gratitude must go to the Faculty of Business Administration and Management of the Universitat Politècnica de València for their unwavering support of CARMA 2024.

2. Organizing Committee

General chair

Josep Domènech, Universitat Politècnica de València

Scientific committee chairs

María Rosalía Vicente, Universidad de Oviedo
Pablo de Pedraza, Universitat Politècnica de València

Steering committee

Aidan Condron, Central Statistics Office, Ireland
Caterina Liberati, Università di Milano-Bicocca
Giuliano Resce, Università di Molise
Josep Domenech, Universitat Politècnica de València
Juri Marcucci, Banca d'Italia
Lisa Crosato, Università Ca'Foscari
María Olmedilla Fernández, Skema Business School, Paris
María Rosalía Vicente, Universidad de Oviedo
Markus Herrmann, Lidl Analytics
Pablo de Pedraza García, Universitat Politècnica de València
Rocío Martínez Torres, Universidad de Sevilla
Sergio Toral Marín, Universidad de Sevilla

Local organization

Jose Baixauli
Eduardo Cebrián
Roberto Cervelló-Royo
Ana Debón
Ana Garcia-Bernabeu
Xin-Hui Huang
José Carlos Huerta Tur
Ana Pastor Merino
Agapito Emanuele Santangelo
Joan Manuel Valenzuela

3. Sponsors and Supporters

Instituto Valenciano de Investigaciones Económicas (IVIE)
Lidl

Generalitat Valenciana
Universitat Politècnica de València
Facultad de Administración y Dirección de Empresas
Departamento de Economía y Ciencias Sociales

4. Scientific committee

Anton Aasa, University of Tartu, Estonia
Fernando Almeida, Ispgaya & Inesc Tec,
María del Pilar Ángeles, Universidad Nacional Autónoma de México
Nikolaos Askitas, IZA – Institute of Labor Economics (IDSC)
Seymus Baloglu, University of Nevada Las Vegas
Catherine Beaudry, Polytechnique Montréal
Nicola Caravaggio, Unviersità degli Studi del Molise
Ramón Alberto Carrasco González, Universidad Complutense de Madrid
Ernesto Cassetta, Department of Economics and Statistics (DIES), University of Udine,
Elena Catanese, Istat
Eduardo Cebrián, Universidad Europea de Valencia
Merce Crosas, Barcelona Supercomputing Center
Lisa Crosato, Ca' Foscari University of Venice
Antonio De Nicola, ENEA
Giuditta de Prato, EC JRC
Kuangnan Fang, Xiamen University
Juan Fernández de Guevara, Ivie & Universitat de València
Yolanda Gomez, Devstat
Marcos González-Fernández, Universidad de León
Agustín Indaco, Carnegie Mellon University in Qatar
Riadh Ladhari, Laval University
Caterina Liberati, Università di Milano-Bicocca,
Rocío Martínez-Torres, Universidad de Sevilla
Jesús Morán, University of Oviedo
Federico Neri, Deloitte
María Olmedilla, SKEMA Business School – Université Côte d'Azure
Enrique Orduña-Malea, Universitat Politècnica de València
José Luis Ortega, IESA/CSIC
Luca Pappalardo, ISTI-CNR
José Manuel Pavía Miralles, Unversitat de Valencia
Arturo Peralta Martín-Palomino, UCLM, España
Virgilio Perez, University of Valencia
Maria Petrescu, Embry-Riddle Aeronautical University

Index

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17776

Who Enjoys the Lion's Share? Unveiling Sentiment of the Media in Indonesia's Presidential Election Using Large Multi-Language Model

Angga Wahyu Anggoro[1], Ani Tri Wahyuni[2]
[1]School of Computer Science and Informatics, Cardiff University, United Kingdom, [2]Cardiff Business School, Cardiff University, United Kingdom.

How to cite: Anggoro, A. W., Wahyuni, A. T. 2024. Who Enjoys the Lion's Share? Unveiling Sentiment of the Media in Indonesia's Presidential Election Using Large Multi-Language Model. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17776

Abstract

The media serves as a framing tool to influence people's decisions. With more than 220 million internet users in Indonesia, it is a promising avenue for gaining significant voter support in the presidential election. This study aims to contribute to existing knowledge by examining how media coverage shapes public perceptions of the presidential election. The research uses a large multi-language model to predict sentiment analysis for the investigation of media framing and cultivation theories. The study examines 778 headlines from 119 media outlets collected from news aggregator services in the campaign period to understand media behaviour throughout the presidential campaign. The findings reveal the alignment between media coverage during the presidential election and the result of the quick count.

Keywords: *Headline News, Sentiment Analysis, XLM-T, Politics, Presidential Election*

1. Introduction

On 14 February 2024, Indonesia held its five-yearly presidential election. According to the General Elections Commission's quick count, the Prabowo-Gibran ticket emerged victorious (Komisi Pemilihan Umum, 2024). Many factors contribute to the outcome of an election, including the media. It is undeniable that the media has had a significant impact on politics (Aririguzoh, 2021). This research seeks to understand media coverage's role in informing voters by investigating the sentiment of news articles published during the presidential campaign. The research seeks to answer the question: What is the media behaviour throughout the presidential campaign in Indonesia?

Studying media coverage of presidential candidates in Indonesia is increasingly valuable due to the country's status as one of the largest electorates with extensive internet exposure (ranking

fourth globally with 224 million internet users in 2022 and an average of 7 hours and 42 minutes daily usage on various media platforms and devices (Statista, 2024)). Given the low literacy levels, susceptibility to manipulation and misinformation among Indonesian citizens (Devega, 2017), it is pertinent to apply media framing and cultivation theory through sentiment analysis of headline news to understand media behaviour during the election process, as certain candidates can use the media for propaganda purposes. Using headline news instead of user opinions will have a greater impact because news provides reliable information for all users. Even social media users often rely on it to form their opinions. Using a large multi-language model (XLM-T), this study examines national and international media headline sentiments over an eight-month period in Indonesian and English. Analysing media framing and cultivation theories, particularly through examining news headline sentiment analysis, remains underexplored as most of the previous research has relied on data from platforms such as Twitter and Facebook. This study seeks to address this gap in the literature and contribute to understanding the media-politics nexus exploits a multilingual pre-trained model to generate sentiment from news headlines.

This study is organised as follows: It begins with an explanation of the related works, followed by a description of the methodology employed. The analysis and discussion are then presented. Finally, the paper concludes with the study's limitations and suggestions for future research.

2. Related work

2.1. Theories in Media Study and Electoral Condition in Indonesia

The existing mutually beneficial connection between online enthusiasm and grassroots political mobilisation reinforces the role of the media in the political sphere. Media content influences public opinion, though not always in direct or precise ways, whether through processes such as "priming" (Iyengar and Kinder, 1987), agenda setting, or the shaping of public debate through news formats (Altheide, 1997).

Framing and cultivation theories are pivotal to understanding the role of media in political movements. According to Klein & Amis (2021), framing theory delves into rhetorical strategies that influence individuals to adopt a particular perspective by strategically emphasising elements of perceived reality. Media use frames to simplify messages, gain support and deter opposition (Klein & Amis, 2021). For example, media portrayals can influence people's decisions by framing information with different emotions and language patterns, especially in headlines. Meanwhile, the cultivation theory by Potter (2014) explores how repeated media exposure shapes perceptions and attitudes over time.

In countries with low literacy rates, such as Indonesia, citizens may struggle to identify the credibility of news sources. Minimal learning habits and extensive screen time exacerbate

susceptibility to misinformation (Devega, 2017). Clickbait headlines also contribute to this trend, often leading to instant interpretation without more profound analysis. Micro-targeting strategies in electoral campaigns exploit these vulnerabilities, influencing voters' opinions and potentially swaying election outcomes (Arugay, 2022).

2.2. Sentiment Analysis with Large Language Models

The growing utilisation of textual data in machine learning has led to increased interest in the natural language processing (NLP) domain, which is applicable in downstream tasks such as sentiment analysis, topic modelling, summarisation, and generative text. Various deep learning architectures take part in this development, including transformer-based models with attention-based architecture (Vaswani et al., 2017), which produce various pre-trained language models. These models, trained on large corpora, generate numerical representations of text sequences, allowing machines to learn syntactic and semantic aspects of words for downstream applications such as sentiment analysis. A model such as XLM RoBERTa (Conneau et al., 2020), trained on multilingual data, serves as an effective baseline for real-world multilingual tasks.

Sentiment analysis, also known as opinion mining, categorises text data based on opinion polarity, such as positive, neutral and negative sentiment (Dashtipour et al., 2018). Sentiment analysis approaches have developed from heuristic or rule-based to the recent machine learning model with the availability of an abundance of unstructured data. Model-based methods are preferred because of their model performance and the ability over human interpretation to reduce bias and emotional involvement.

3. Methodology

Our approach in this study is shown in Figure 1. Firstly, we pre-processed the news headlines and supplied the data into the model to obtain the sentiment of news related to the presidential election. An additional tool is then used to explain the model behaviour towards the sentiment predictions visually.

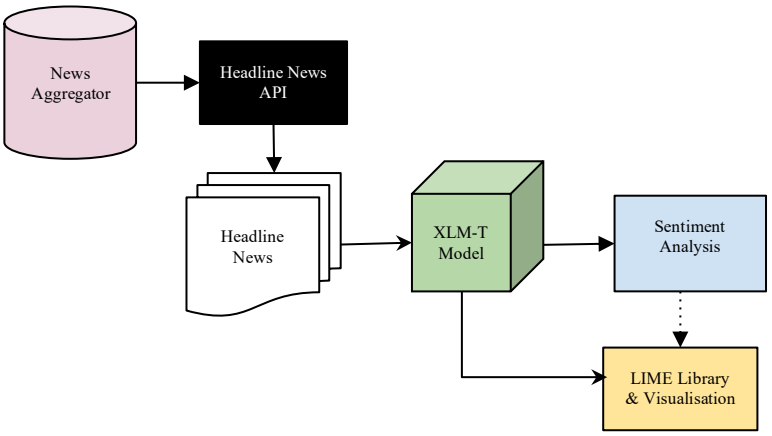


Figure 1. Diagram of Headline News Sentiment Analysis

3.1. Data pre-processing

The datasets used in this research were obtained from a public news aggregator, which aggregates news headlines from various global publishers. We used a publicly available API to obtain articles and retrieve our query in JSON format. Several parameters were used to define the scope of the data collection, including the period of the data, which is from July 2023 to the end of the campaign period. As certain media organisations may show favouritism towards certain candidates, we refrain from cherry-picking news outlets to mitigate bias. The language parameter was left unrestricted, but we anticipated that the results would predominantly be in either Indonesian or English. Several keywords are used for the query, including the full names, nicknames and aliases of relevant candidates, as outlined in Table 1.

Table 1. Name of Presidents and Vice President Candidates

1st Candidates	2nd Candidates	3rd Candidates
Anies	Prabowo	Ganjar
Anies Baswedan	Prabowo Subianto	Ganjar Pranowo
Muhaimin Iskandar	Gibran	Mahfud MD
Cak Imin	Gibran Rakabuming Raka	Prof Mahfud

Some Python libraries were used for data pre-processing, data explanation, and visualisation purposes. Following the data collection, we obtain the JSON-formatted response from API that provides five attributes for each headline: title, description, published date, URL, and publisher. In the pre-processing phase, we eliminate duplicate entries and remove the publisher names from the headline titles before putting the data into the model. We then employ a large multi-language model to generate the sentiment of the headline news.

3.2. Sentiment Analysis with XLM-T model

In this study, we exploit the XLM-T model (Barbieri et al., 2022). This model was trained on millions of tweets from over thirty languages, including English and Indonesian, and it provides zero shots of multilingualism for sentiment analysis tasks. The result competes with XLM-R, the baseline pre-trained for multilingual sentiment analysis. We use the model straightforwardly without any fine-tuning; instead, we inferred our query to generate sentiment from datasets after obtaining the model that can be accessed from HuggingFace[1].

The LIME library (Ribeiro et al., 2016) is employed to improve the interpretability of the model's behaviour, such as large language models (LLMs). It explains how the agnostic model makes predictions by delineating the contributions of features from the input data. It works effectively with tabular, image and text data. By using interpretable models such as linear models or decision trees as surrogate models, LIME approximates the black box model by generating samples around a particular point of interest. For text data, LIME works by turning certain word patches on and off and assessing which text features contribute to the model's prediction.

4. Result and Analysis

The total number of news headlines in datasets is 1094 from 119 media outlets, with a reduction after duplicates and uncorrelated news are eliminated, leaving 778. The analysis shows that all candidates receive media coverage, albeit to varying degrees. Ganjar-Mahfud, for example, receives the least media attention of the candidates. Figure 2 illustrates a spike in news volume towards the end of the campaign, with the third-placed candidate receiving the fewest mentions while the second-placed candidate receiving the most news mentions. This trend is in line with the election result according to quick counts, where the candidate in second place (Prabowo - Gibran) received the most votes, followed by the candidate in first place (Anies - Muhaimin) and then the candidate in third place (Ganjar - Mahfud) (Komisi Pemilihan Umum, 2024).

[1] *https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment*

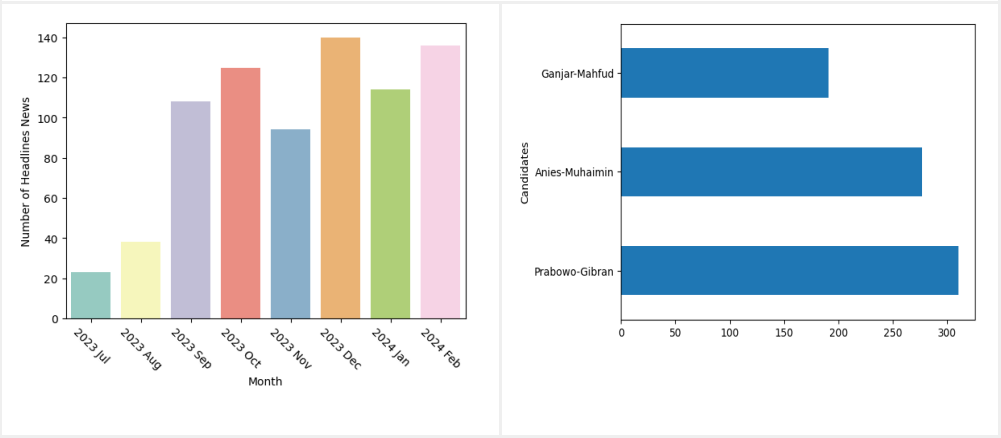


Figure 2. The Statistics of Headlines News

Figure 3 depicts the sentiment analysis generated by XLM-T models, which are covered by the media towards all candidates and have the highest occurrence of neutral sentiments. The news headlines for both the second and third candidates show a consistent sentiment pattern, with a lower frequency of negative sentiment compared to positive sentiment. In contrast, the first candidate receives a higher frequency of negative sentiment than positive ones.

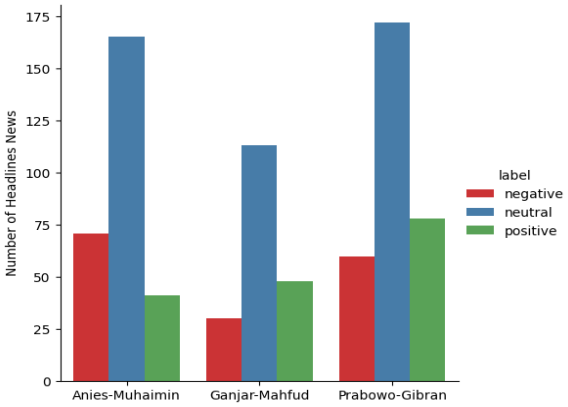


Figure 3. The Sentiment of Headlines News

The sentiment analysis results indicate that the sentiments conveyed (framing) and the repeated exposure (cultivating) by the media are parallel with the outcome of the Indonesian democratic process, as evidenced by the proportional vote results for each candidate in the quick count.

Figure 4 illustrates examples from the LIME package that explain the behaviour of the XLM-T model in predicting headline sentiment. The word 'maintains' contributes the highest positive score, classifying the first headline as positive. In contrast, 'bloody' generates the highest negative score for the second headline, resulting in an overall negative sentiment classification.

This method demonstrates the model's word weighting and sentiment classification and provides insight into the media's choice of words to portray the presidential candidates.

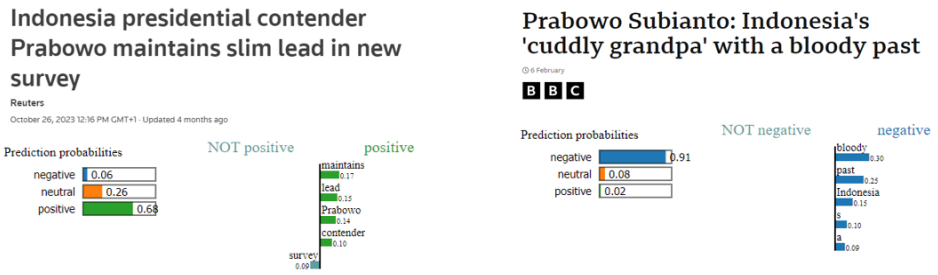


Figure 4. LIME sentiment explanation for the headline news

LIME provides explainability for individual predictions, giving ordinary users confidence in using a reliable model for downstream applications. With a reliable multilingual model, users, or in this case, the succession team, can predict candidate sentiment in local or foreign languages for headline news and mitigate negative sentiment for advocacy because continued negative influence may lead to scepticism among swing voters. The successor team can also assess news outlets to identify those that frequently frame a particular candidate with a certain sentiment.

5. Limitation

While most of the headlines align with our expectations, there is some ambiguity. For example, when two candidates are mentioned, it becomes difficult for the model to assign sentiment. In addition, the model struggles to capture contextual nuances as it relies on the connotations of words to determine sentiment. Given the brevity of headlines, it is difficult to capture the sentiment accurately. Furthermore, we acknowledge limitations such as using the model without fine-tuning our datasets and lacking ground truth certainty, which complicates analysis.

6. Conclusion

In this study, we analyse news headlines to investigate sentiments surrounding the Indonesian presidential election using framing and cultivation theories. Using the XLM-T model, we predict sentiments in collected headlines for each candidate and explain them visually using LIME. Our results show that media coverage aligns with the election outcome, with the winner receiving the most coverage and significant positive sentiment. This model could be useful to campaign teams or consultants in anticipating news impact and maintaining candidate tone in the media. Given the limitations, future research incorporating public participants and language experts is recommended for accurate sentiment analysis.

Acknowledgement

The authors would like to thank the Indonesia Endowment Fund for Education (LPDP), Ministry of Finance Indonesia for supporting this research.

Reference

Altheide, D. L. (1997). "The News Media, the Problem Frame, and the Production of Fear". The Sociological Quarterly, 38(4), 647–668. https://doi.org/10.1111/j.1533-8525.1997.tb00758.x

Aririguzoh, S. A. (Ed.). (2021). Global Perspectives on the Impact of Mass Media on Electoral Processes: IGI Global. https://doi.org/10.4018/978-1-7998-4820-2

Arugay, A. A. (2022). Stronger Social Media Influence in the 2022 Philippine Elections. Fulcrum, Analysis of South East Asia. https://fulcrum.sg/stronger-social-media-influence-in-the-2022-philippine-elections/

Barbieri, F., Espinosa Anke, L., & Camacho-Collados, J. (2022). XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 258–266). European Language Resources Association.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 8440–8451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

Dashtipour, K., Hussain, A., & Gelbukh, A. (2018). Adaptation of Sentiment Analysis Techniques to Persian Language. In A. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing (Vol. 10762, pp. 129–140). Springer International Publishing. https://doi.org/10.1007/978-3-319-77116-8_10

Devega, E. (2017). Teknologi masyarakat Indonesia: Malas baca tapi cerewet di medsos. Kominfo, Indonesia Terkoneksi. https://www.kominfo.go.id/content/detail/10862/teknologi-masyarakat-indonesia-malas-baca-tapi-cerewet-di-medsos/0/sorotan_media

Iyengar, S., & Kinder, D. R. (1987). News That Matters: Television and American Opinion. The University of Chicago Press.

Klein, J., & Amis, J. M. (2021). The Dynamics of Framing: Image, Emotion, and the European Migration Crisis. Academy of Management Journal, 64(5), 1324–1354. https://doi.org/10.5465/amj.2017.0510

Komisi Pemilihan Umum. (2024). Info Publik Pemilu 2024. https://pemilu2024.kpu.go.id/

Potter, W. J. (2014). A Critical Analysis of Cultivation Theory: Cultivation. Journal of Communication, 64(6), 1015–1036. https://doi.org/10.1111/jcom.12128

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why Should I Trust You?': Explaining the Predictions of Any Classifier (arXiv:1602.04938). arXiv. https://doi.org/10.48550/arXiv.1602.04938

Statista. (2024). Indonesia: Number of internet users 2028. https://www.statista.com/statistics/254456/number-of-internet-users-in-indonesia/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. Nips. http://arxiv.org/abs/1706.03762

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17783

Economic forecasting with non-specific Google Trends sentiments: Insights from US Data

Sami Diaf[1], Florian Schütze[2]

[1]Department of Socioeconomics, University of Hamburg, Germany. [2]Faculty of Economics and Social Sciences, Helmut Schmidt University, Germany.

How to cite: Diaf, S.; Schütze, F. 2024. Economic forecasting with non-specific Google Trends sentiments: Insights from US Data. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17783

Abstract

The influence of specific Google Trends search queries measuring various sentiments on economic performance and stock markets has been extensively documented and used for many purposes. This paper examines the predictive power of queries measuring non-specific sentiment on key macroeconomic variables when linked to a comprehensive sentiment dictionary. The analysis shows that non-specific sentiments do not improve the forecasting quality of the US economy as a whole, except for unemployment, which was found to be predictable for all sentiments. Consequently, the authors suggest that economic-related sentiments with carefully selected words should be used in Google Trends search queries to improve predictive performance. However, if a socio-cultural analysis is to be performed, non-specific sentiments would be suggested, as they can be predicted by the real economic time series of unemployment.

Keywords: *Sentiment analysis; Google Trends and Search Engine data; Web scraping; Internet econometrics; Forecasting and nowcasting.*

1. Introduction

Social listening is a regular source of data used by economists to test further hypotheses beyond the available economic aggregates and other indices. The sentiment of economic agents could be used as a proxy to predict the evolution of economic time series.

Affuso & Lahtinen (2019) showed that negative sentiment among Twitter users has a greater impact on stock returns than positive sentiment. Eugster & Uhl (2024) established an improvement in the accuracy of an inflation forecast using a self-generated sentiment index based on newspaper articles. Rambaccussing & Kwiatkowski (2020) also used sentiment analysis of newspaper articles and found it useful for forecasting unemployment and output. Sharpe et al. (2017) analyzed optimistic and pessimistic sentiment in Federal Reserve Board forecasts and discovered that these sentiments can predict both GDP growth and unemployment.

Since the above literature has shown that sentiment measured in various media has an impact on the real economy, this paper will investigate whether this can be applied to Google Trends and non-specific sentiment. There are examples in the literature that this works with specific sentiment words.

Broachado (2020) created a Google Sentiment Index that measures the overall polarity about the economy and shows short-term predictive ability regarding the stock market. Borup & Schütte (2022) show that labor market forecasts can be improved by using specific labor market searches on Google. While these papers use Google Trends with keywords, it has also been shown that economic uncertainty sentiment, measured with economic topics instead of keywords, has an impact on the economy (Schütze, 2020; Schütze, 2022). Donadelli and Gerotto (2019) showed that an increase in search queries for non-macro-based topics had a negative impact on economic time series. However, these topics were related to health, environment, security, and politics, which means that they are also specific in some sense.

Therefore, this paper explores the possibility of improving the forecasting quality of economic time series by analyzing non-specific sentiments in Google Trends, with the goal of improving the forecasting quality of macroeconomic time series. For this purpose, the sentiment dictionary of Loughran & McDonald (2011), which contains 8 different sentiment categories, is used, although its scope is not limited to the economic context, but the words measure the basic sentiment in a publication or in Google Trends. Thus, this application investigates the duality of specific/non-specific sentiments and their predictive abilities in the economic domain.

The results show that non-specific sentiments do not increase the overall explanatory power of macroeconomic variables, but only two out of eight sentiments have an impact on some of the economic time series. This shows that the approach with specific sentiment words is justified, as Google Trends time series with non-specific words do not have the same predictive quality as specific words. Moreover, unemployment is the only aggregate that has a predictive quality for all 8 sentiments, meaning that it captures the interest of most economic agents as translated by their search queries.

2. Methodology

Loughran & McDonald's (2011) sentiment word list has been used for many sentiment analysis exercises in the social sciences and, more recently, on social listening sources (Google Trends). It contains eight (8) sentiment categories: Negative, Positive, Uncertainty, Litigious, Strong_Modal, Weak_Modal, Constraining, and Complexity. Table 1 shows the different sentiment categories and the number of words within each category. It also shows the number of words that resulted in a complete time series of Google Trends search queries. Approximately 10% of the words did not result in a complete Google Trends query.

For each word in the sentiment word list, a Google Trends query was run using the R package "gtrendsR" (Massicotte et al., 2016). A total of 4,194 queries were run, of which 3,740 resulted in a Google Trends time series. This means that for each of the 3740 different queries, a Google Trends time series for the US was downloaded for the period 01/2004 to 12/2023.

A total of eight Google Trends sentiment indices were created from the words in each sentiment category. Four different "weights" were applied: 1. The average of all Google Trends time series within a category; 2. The weighted average of all Google Trends time series. The weight is determined by the relative frequency with which each word occurs within a category (Loughran & McDonald, 2011); 3. A principal component analysis was performed on all words in each category. The third weight is the first principal component of the analysis, and the fourth weight is the second principal component. The first explains approximately 40% of the data, the second approximately 20% in each sentiment category.

Table 1. The number of words in the sentiment dictionary from Loughran & McDonald (2011) and the number of Google queries available to match them. Source: Own calculation.

	Negative	Positive	Uncertainty	Litigious	Strong_Modal	Weak_Modal	Constraining	Complexity	Sum
Words	2355	354	297	905	19	27	184	53	4194
Queries	2195	289	294	727	17	27	149	42	3740

Figure 1 shows the eight different sentiment categories and the four different weightings. Note that positive and negative sentiment show a similar trend in the two different weightings. Only the principal component analysis shows an opposite trend for negative and positive sentiment. The same is true for the distinction between Weak_Modal and Strong_Modal sentiment. The chart shows that the approach of testing different weightings can be useful, as there are significant differences.

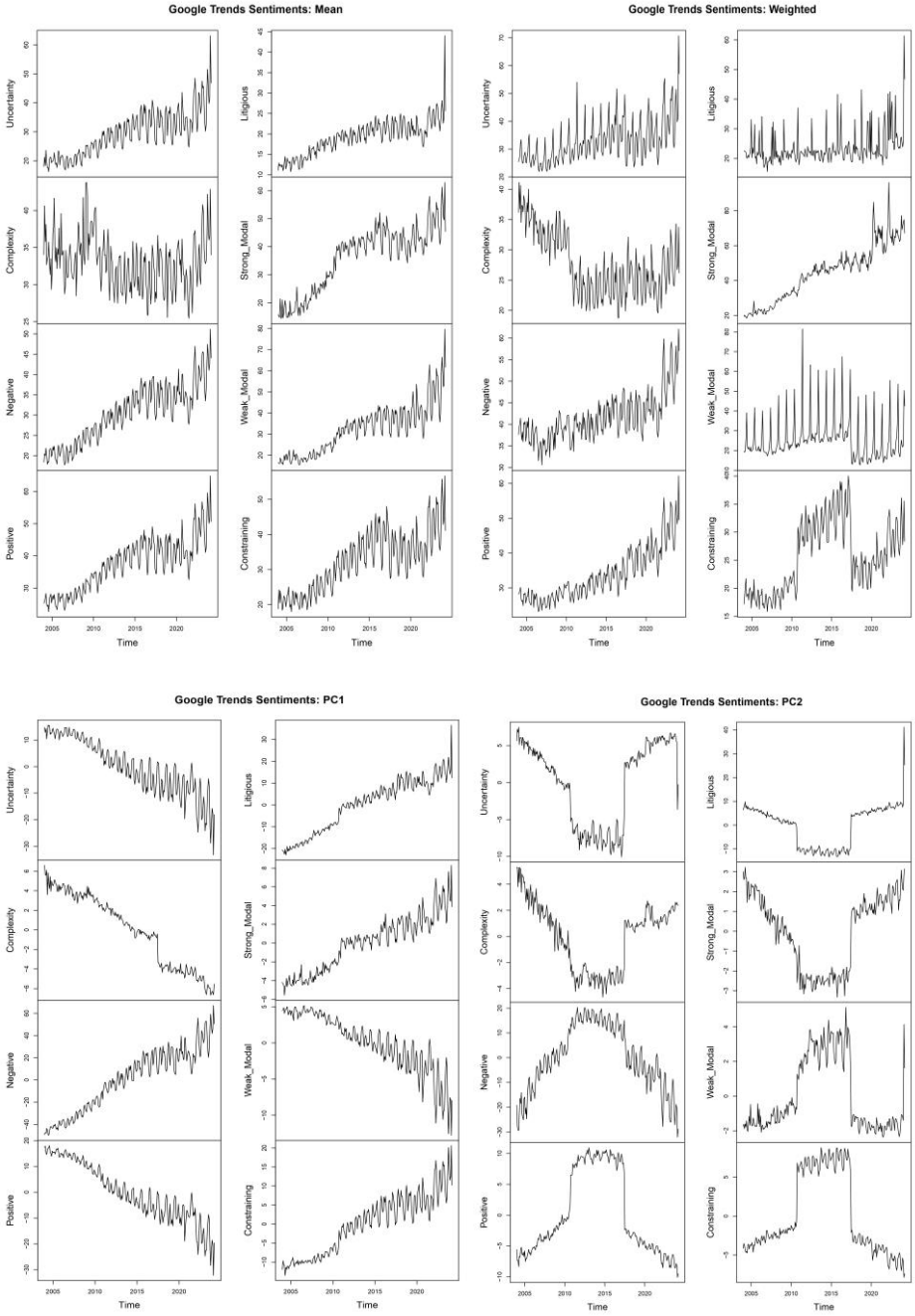


Figure 1. Google Trends Sentiments: different weightings. Source: Own calculation.

The Granger causality analysis used all 8 different sentiments with the 4 different weights, resulting in 32 time series, as well as 4 economic monthly time series: personal consumption expenditures (U.S. Bureau of Economic Analysis, 2024), consumer prices (OECD, 2024b), monthly unemployment rates (OECD, 2024c), and industrial production (OECD, 2024a). The time period is from 01/2004 to 12/2023. The research question to be tested here is whether the underlying sentiment, as measured by Google Trends, has predictive power for real economic time series. At the same time, it is also tested whether the real economic time series can predict the sentiment. It is important to note that Granger causality is not causality in the classical sense, but only shows that one time series has predictive power with respect to another time series.

All series are seasonally adjusted by creating a dummy variable for each month. The seasonally adjusted series is the residual series from a regression with the original series as the endogenous variable and the months without constants as the exogenous variable. The Augmented Dickey-Fuller test was then used to determine whether the time series were I(0) or I(1). The method of Toda & Yamamoto (1995) was used for the Granger causality analysis because it is robust to different stationary orders of the time series. Since different time series are analyzed, the following combination could occur: Sentiment time series = I(0) and industrial production= I(1).

The lag selection for the VAR model was done using Akaike's information criterion, with a maximum lag length of 12. A time series is assumed to be Granger causal for another time series if the null hypothesis that there is no Granger causality can be rejected at the 5% significance level. The Wald test used in Granger causality analysis is based on the rather simple premise of comparing the performance of a restricted model Y, which excludes X, with an unrestricted model for Y, which includes X.

3. Results

The analysis of Granger causality in the direction of sentiment towards the economic time series in Table 2 shows that sentiment does not have much predictive power. The null hypothesis that sentiment Granger-causes consumer prices cannot be rejected with $\alpha<5\%$ for a single sentiment time series. Industrial production can only be Granger-caused by two sentiment time series, the first principal component of the positive sentiment category and the weighted Strong_Modal sentiment. Personal consumption expenditures can be Granger-caused by 3 sentiment time series: The average of the Litigious sentiment, the first principal component of the Litigious sentiment, and the average of the Strong_Modal sentiment. Unemployment can also be Granger causally explained by three sentiment time series. In this case, it is the first principal component of the Complexity Sentiment. In addition, the average of the Litigious Sentiment and the average of the Strong_Modal Sentiment can Granger causally explain unemployment.

Therefore, there are sentiments that in some cases have increased predictive performance. These include the Strong_Modal and Litigious categories. Given that 32 different sentiment time series

were used, with different weights for each economic variable, it is expected that 5% of the 32 models will reject H0 of the Granger test by chance. On average, this would be 1.6 models. On average, 2 models per economic variable lead to a rejection of H0. Given this, the authors conservatively assume that Google Trends sentiment has no fundamental Granger causality on economic time series.

Table 2. p-values of the Granger causality analysis, both directions. In red and bold: α<5% when H0 is rejected. Source: Own calculation.

	H0: Sentiment does not Granger cause one of the economic time series					**H0: An economic time serie does not Granger cause one of the sentiments**			
	CPI	**Ind. Prod.**	**PCE**	**Unemp.**		**CPI**	**Ind. Prod.**	**PCE**	**Unemp.**
UncertaintyMean	0,890	0,455	0,849	0,562		0,547	0,762	0,096	**0,006**
UncertaintyWeight	1,000	0,672	0,665	0,476		0,852	0,328	0,129	0,107
UncertaintyPC1	0,913	0,239	0,574	0,309		0,688	0,269	**0,041**	**0,000**
UncertaintyPC2	0,957	0,132	0,938	0,112		0,929	0,814	0,999	0,088
ComplexityMean	0,519	0,737	0,816	0,535		0,099	0,059	0,464	**0,003**
ComplexityWeight	0,986	0,436	0,831	0,485		0,162	0,242	0,103	**0,000**
ComplexityPC1	0,688	0,709	0,876	**0,030**		0,214	0,932	0,805	0,136
ComplexityPC2	0,108	0,614	0,968	0,275		0,389	0,802	0,730	0,154
NegativeMean	0,206	0,470	0,067	0,160		0,535	0,431	0,358	0,117
NegativeWeight	0,753	0,849	0,341	0,500		0,103	0,803	0,888	0,632
NegativePC1	0,103	0,484	0,052	0,126		0,865	0,231	0,132	**0,022**
NegativePC2	0,963	0,371	0,237	0,148		0,788	0,552	0,233	0,051
PositiveMean	0,421	0,081	0,155	0,178		0,136	0,198	0,174	**0,027**
PositiveWeight	0,913	0,102	0,192	0,091		0,456	0,254	0,536	0,056
PositivePC1	0,489	**0,040**	0,232	0,163		0,285	0,130	0,304	**0,006**
PositivePC2	0,551	0,739	0,872	0,299		0,836	0,501	0,866	**0,027**
LitigiousMean	0,101	0,152	**0,017**	**0,033**		0,362	0,516	0,388	0,473
LitigiousWeight	0,298	0,366	0,136	0,367		**0,023**	0,563	0,661	**0,004**
LitigiousPC1	0,207	0,186	**0,028**	0,106		0,843	0,724	0,449	0,214
LitigiousPC2	0,466	0,564	0,712	0,201		0,783	0,849	0,472	0,120
Strong_ModalMean	0,328	0,080	0,846	0,608		0,652	0,760	0,748	**0,006**
Strong_ModalWeight	0,172	**0,000**	**0,023**	**0,000**		0,883	0,340	**0,033**	0,383
Strong_ModalPC1	0,693	0,187	0,674	0,418		0,858	0,923	0,778	**0,030**
Strong_ModalPC2	0,556	0,885	0,943	0,488		0,732	0,236	0,472	0,244
Weak_ModalMean	0,941	0,572	0,403	0,424		0,601	0,701	0,296	**0,014**
Weak_ModalWeight	0,961	0,922	0,928	0,844		0,777	0,964	0,997	0,772
Weak_ModalPC1	0,970	0,510	0,330	0,368		0,127	0,264	0,166	**0,001**
Weak_ModalPC2	0,628	0,995	1,000	0,722		0,769	0,908	0,994	0,303
ConstrainingMean	0,745	0,393	0,620	0,535		0,408	0,662	0,116	**0,007**
ConstrainingWeight	0,996	0,493	0,964	0,657		0,454	0,722	0,561	0,251
ConstrainingPC1	0,900	0,373	0,442	0,288		0,624	0,826	0,143	**0,008**
ConstrainingPC2	0,922	0,388	0,990	0,302		0,977	0,999	0,961	0,077

When the Granger causality sequence is reversed, it turns out that consumer prices can only Granger causally explain the average of the litigious. Industrial production cannot Granger

causally explain a single sentiment. Consumer spending has predictive power for the first principal component of uncertainty. In addition, consumer spending can explain the average of Strong_Modal. Unemployment can Granger causally explain a total of 15 out of 32 sentiments. In each sentiment category, there is at least one weighting scheme that is Granger causally influenced by unemployment.

Contrary to the previous statement that Google Trends sentiment has no predictive quality for economic variables, it is now apparent that US unemployment has predictive quality for Google Trends sentiment. With this discovery, the focus of this paper changes from a (macro)economic analysis to a sociological analysis. It turns out that general, non-economic sentiments have no influence on future economic development. On the other hand, a change in unemployment affects all categories of sentiment and these can be better predicted. It seems that the socio-cultural influence of unemployment on the mood of the population is very strong and pronounced.

4. Conclusion

The initial working hypothesis that unspecific sentiment, as measured by Google Trends, can be used to improve the forecasting performance of monthly economic time series is likely to be rejected. It turns out that only two sentiments show improved forecasting performance, while the remaining do not. In contrast, a reverse Granger causality analysis shows that unemployment has an impact on each sentiment category. This is confirmed by different weighting schemes. If this predictive quality is equated with an influence from one time series to another, it becomes clear that unemployment has a strong influence on the polarization of society. Even without this equation, it is obvious that there must be a correlation, since lack of work can be an existential experience that also strongly polarizes individuals.

Furthermore, this paper shows that the results of previous applications of sentiment measurement cannot be generalized. For example, non-specific sentiment uncertainty does not affect economic variables, but specific economic sentiment "uncertainty" does.

The finding that unemployment has a Granger causal effect on all sentiments shows that the Google Trends approach is promising. Theoretically, a change in unemployment should capture the interest of society, which in turn should lead to more Google searches. However, this finding also reinforces the statement that only (economic) specific sentiments should be used if the predictive power of these sentiments is to be increased. This can be done by choosing an appropriate topic or by concatenating words. Another area of research would be to investigate the specific influence of unemployment on sentiments, and whether this is the case across regions. For example, urban regions with an affinity for the Internet may be more affected than suburban or rural regions.

References

Affuso, E., Lahtinen, K.D. (2019). Social media sentiment and market behavior. Empirical Economics 57, 105–127.

Borup, D., & Schütte, E. C. M. (2022). In search of a job: Forecasting employment growth using Google Trends. Journal of Business & Economic Statistics, 40(1), 186-200.

Brochado, A. (2020). Google search based sentiment indexes. IIMB Management Review, 32(3), 325-335.

Donadelli, M., and Gerotto, L. (2019). Non-macro-based Google searches, uncertainty, and real economic activity. Research in International Business and Finance, 48, 111-142.

Eugster, P., and Uhl, M. W. (2024). Forecasting inflation using sentiment. Economics Letters Economics Letters, Volume 236, 111575.

Massicotte, P., Eddelbuettel, D. and Massicotte, M. P. (2016). Package 'gtrendsR', R package.

Loughran, T. and McDonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance, 66 (1), 35-65.

OECD (2024a). Industrial production (indicator), doi: 10.1787/39121c55-en (Accessed on 14 February 2024).

OECD (2024b). Inflation (CPI) (indicator), doi: 10.1787/eee82e6e-en (Accessed on 14 February 2024)

OECD (2024c), Unemployment rate (indicator), doi: 10.1787/52570002-en (Accessed on 14 February 2024)

Rambaccussing, D. and Kwiatkowski, A. (2020). Forecasting with news sentiment: Evidence with UK newspapers. International Journal of Forecasting, 36 (4), 1501-1516.

Schütze, F. (2020), Google Trends Topic-Based Uncertainty: A Multi-National Approach, CARMA 2020 - 3rd International Conference on Advanced Research Methods and Analytics, Valencia, Spain, 8 – 9 July 2020, pp. 191-199.

Schütze, F. (2022), The demand side of information provision: Using multivariate time series clustering to construct multinational uncertainty proxies, CARMA 2022 - 4rd International Conference on Advanced Research Methods and Analytics, Valencia, Spain, 29 June – 1 July 2022, pp. 155–163.

Sharpe, S A., Sinha, N. R., and Hollrah, C. A. (2017). What's the Story? A New Perspective on the Value of Economic Forecasts. Finance and Economics Discussion Series 2017-107, Board of Governors of the Federal Reserve System (U.S.).

Toda, H. Y. and Yamamoto, T. (1995). Statistical inference in vector autoregressions with possibly integrated processes. Journal of Econometrics, 66 (1-2), 225–250.

U.S. Bureau of Economic Analysis (2024). Personal Consumption Expenditures [PCE], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/PCE (Accessed on 14 February 2024).

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17805

A multi-platform framework for nowcasting social phenomena: a case study for food insecurity

Bia Carneiro[1], Giuliano Resce[2], Giulia Tucci[3], Giosuè Ruscica[2], Nicola Caravaggio[2], Laura Fanelli[2], Agapito Emanuele Santangelo[2], Pietro Cruciata[2]

[1]Bioversity International, Italy, [2]Department of Economics, University of Molise, Italy, [3]International Center for Tropical Agriculture, Brazil.

How to cite: Carneiro, B.; Resce, G.; Tucci, G.; Ruscica, G.; Caravaggio, N.; Fanelli, L.; Santangelo, A. E.; Cruciata, P. 2024. A multi-platform framework for nowcasting social phenomena: a case study for food insecurity. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17805

Abstract

Given the growing significance of internet-based information flows, this research proposes a conceptual framework that integrates digital platforms to nowcast social phenomena, applied to the context of food security monitoring in the Global South. Building on the foundations of Digital Methods and online issue mapping, our research objective is to establish a multi-modal, multi-media model that monitors events from different perspectives to identify potential early warning signals arising from the data, ultimately informing policy actors and supporting early action. We apply three analytical processes: social listening, media monitoring and search interest analysis. Exploratory analysis on data from Zimbabwe point to the feasibility of the models applied to identify food security dimensions in text and search engine data. Further analysis is needed to interpret converging and diverging trends across the data streams, and their implications to food insecurity early warning.

Keywords: *digital platforms; digital methods; nowcasting; food security*

1. Introduction

Given the growing significance of internet-based information flows (Carneiro et al., 2022), this research proposes a conceptual framework that integrates digital platforms to nowcast social phenomena, applied to the context of food security monitoring in the Global South. The issue is pertinent because despite years of decrease, there is a current upsurge in food insecurity, and addressing this challenge requires comprehending the various factors contributing to it (Balashankar et al., 2023; Queiroz et al., 2021). However, while hazards to food security like climate shocks, conflicts, or market disturbances are extensively monitored through traditional

quantitative and qualitative means, the application of digital platforms analysis can enhance localized knowledge about potential areas of concern.

2. A framework to nowcast food insecurity

Figure 1 presents a systematization of our proposed framework. Our research objective is to establish a multi-modal, multi-media model that monitors events from different perspectives to identify potential early warning signals arising from the data, ultimately informing policy actors and supporting early action. The framework relies on the foundations of the digital methods epistemology, which seeks to explain social phenomena through online dynamics (Rogers, 2013) and specifically on the approach of online issue mapping (Rogers, 2015). Online issue mapping brings forth a set of digital techniques for the detection, analysis and visualisation of topical affairs. It leverages on textual, visual and network analysis to understand how issue are formed in digital environments. Its interdisciplinary approach aims to answer questions such as 'is this topic an issue?', 'who are its actors?' 'what are its animating concerns?', and 'where are the issues happening (media, institutional locations, geography)?'.

It has been recognized that online spaces offer opportunities for analysing and visualising contemporary issue dynamics due to several aspects: issue traces are accessible online; the analytics leverage on the dynamic features and affordances of online media; and digital platforms provide data (for instance, metadata, links, hashtags, mentions, etc.) that can be structured for systematic analysis (Carneiro et al., 2022).

To achieve our aim, we apply three analytical processes. First, social listening and media monitoring aim to cover different dimensions of public discourse, at different timespans. While the social media dimension provides insights into the interests, concerns, and opinions of the general public, news media enables insights into "traditional" event coverage, with more established actors and discourses. Thirdly, as the Google search engine is the most visited website in the world, understanding what people are searching in it can reveal the level of interest in a particular topic. Effectively, the predictive potential of Google Trends supports social listening and media monitoring for nowcasting, as evidenced in the literature (Choi & Varian, 2012). The next section describes the final levels of the framework, pertaining to platforms and data sources.

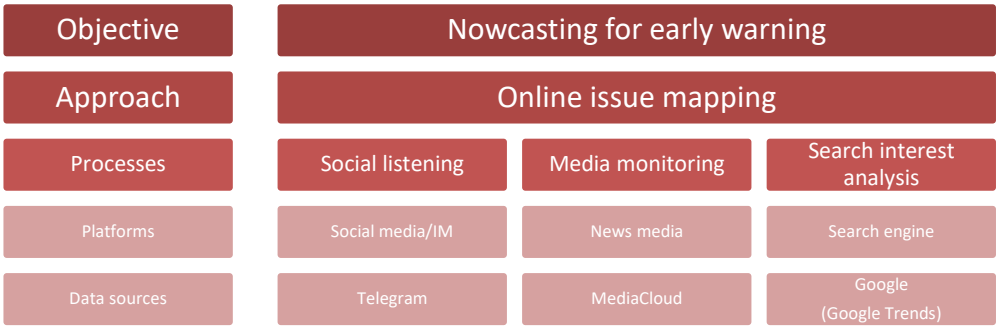


Figure 1. A multi-platform framework to monitor strategic issues

3. Data and methods

3.1 Data sources

For the social listening component, Telegram has been selected. Launched as a messaging app in 2013, Telegram is a currently a digital platform that allows users to create and subscribe to broadcasting channels and create and join discussion groups. It currently has 800 million active users on average, per month. Most importantly, Telegram provides an open API, which emables the extraction of publicly available data from channels and groups.

Media monitoring is performed through data provided by Media Cloud, an open source platform for media analysis aimed at academic research[1]. The platform monitors more than 60 thousand online media sources, with stories processed daily. It also contains curated country collections, enabling granularity in local news collection.

Lastly, search engine data is collected though Google Trends, adapting the daily economic sentiment index (DESI) approach proposed by Eichenauer et al. (2022).

3.2. Data collection

A dedicated software was developed to extract, filter, and visualize Telegram data available on the platform's API. The data collection begins with the creation of curated lists of relevant Telegram groups and channels, which are uploaded to the tool and their content can then be queried, exported and manipulated.

[1] https://www.mediacloud.org

To capture diverging perspectives, data collection covers content generated within our countries of interest (i.e. based on groups and channels from the country in question), as well as content that mentions the country in external or global groups.

News media is queried through Media Cloud's API following a similar structure, where national and local news sources from the countries in question are collected, as well as news stories from global sources that mention the country in the headline.

For Google Trends, the first step of data collection involved determining the main predictors of food insecurity. For this, we leveraged in prior work that used natural language processing to classify food security dimensions in publicly available reports by the USAID's Famine Early Warning Systems Network (FEWS NET). A LASSO regression was applied on the classification results from the NLP model to identify the predictors of food insecurity . The top ten positive features were selected, after excluding those disease-related, such as Covid-19 or Ebola, due to their time or geographic specificity. For each country of interest, a long-run frequency-consistent daily trend of food insecurity was constructed.

3.3. Data analysis

Each of our sources require analytical approaches that leverage on their affordances. Text mining models developed for topic classification and sentiment analysis of food security reporting by FEWS NET were adjusted and combined with visual analysis and machine learning models to monitor and identify the prevalence of food insecurity drivers from diverse perspectives and timescales. Determining the continuous associations and dynamics between drivers supports the identification of potential food insecurity hotspots.

Supervised text mining is applied to all textual data using the previously developed analytical framework and taxonomy for detection of food security-related topics. Leveraging on visual media disseminated on Telegram, the BLIP algorithm (Junnan Li et al., 2022) is applied to generate image captions that describe any images shared on posts. These captions are then combined with the text from the post body, and topic classification is applied. In addition, sentiment analysis is performed using the Syuzhet package (Jockers, 2015). TF-IDF analysis is also performed to identify significant and emerging terminology. Specifically for news media, existing Media Cloud algorithms for entity recognition are used to detect people, organizations, and geographic coverage of news stories.

For search interest, the final time series constructed for each country represents a synthetic search interest (SSI) index for food insecurity based on Google Trends data. In the final step of our procedure, weights were applied to allow for cross-country comparison. The first weight has been constructed by considering the worldwide Google Trends interest of each topic while the second one relied on yearly data of internet penetration (WB, 2024) (inverse) to compensate for the digital divide among countries.

4. Case study: Preliminary results from Zimbabwe

Zimbabwe was selected as the pilot country to test the application of this framework. The results presented here are descriptive in nature as the analysis is ongoing and further interpretation is needed.

670,562 news stories were collected through Media Cloud 2018-2023. Text mining was applied to the headlines. Figure 2 shows the most prevalent topics in news stories from Zimbabwe, and figure 3 presents the aggregated sentiment for these stories.

A list of Telegram groups and channels was curated, as well as a list of Africa-level groups and channels. Four lists of Zimbabwean national and local news sources were used for news data collection, as well as news stories from global English language news sources that mention Zimbabwe. In total, 113,279 posts were collected between 2017-2023. Figure 4 shows the most prevalent topics in groups and channels from Zimbabwe, and figure 5 presents the aggregated sentiment for these posts. The synthetic search interest index (SSI) for food insecurity in Zimbabwe is presented in figure 6.

Creating a nowcasting tool to detect food insecurity and food shocks requires the integration of the diverse data streams (as represented in Figures 2-6) into a unified analysis platform. Time-series analysis enables to track trends over time, employing models that can handle seasonality, trends, and irregular patterns to predict imminent shocks or stressors. Anomaly detection algorithms can be applied to identify sudden changes in the data that deviate from historical patterns, which could indicate emerging crises. Real-time changes in the data can be visualized through an integrative dashboard.

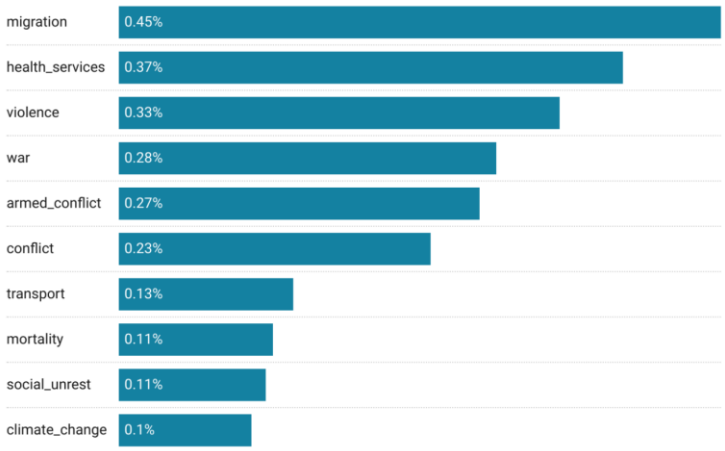


Figure 2. Average percent of headline words dedicated to each topic. Covid-19 has been excluded.

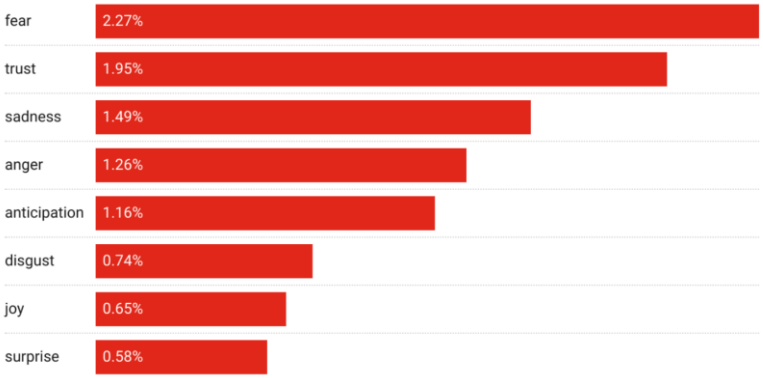


Figure 3. Percent of headlines that registered each emotion.

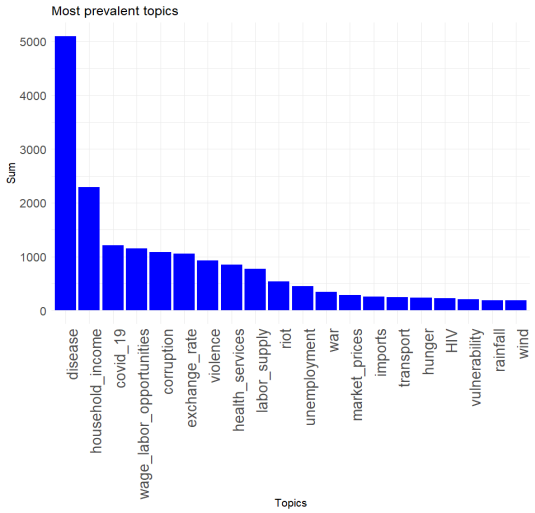


Figure 4. 20 most prevalent topics in Telegram groups and channels from Zimbabwe. Groups related to cryptocurrencies and "buy and sell" groups have been excluded.

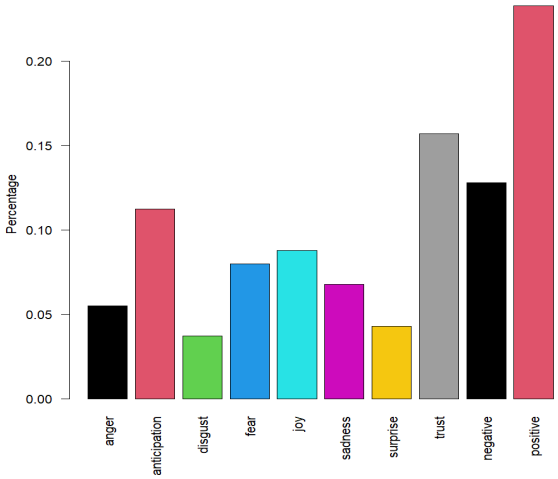


Figure 5. Aggregated sentiment in Telegram groups and channels from Zimbabwe. Groups related to cryptocurrencies and "buy and sell" groups have been excluded.

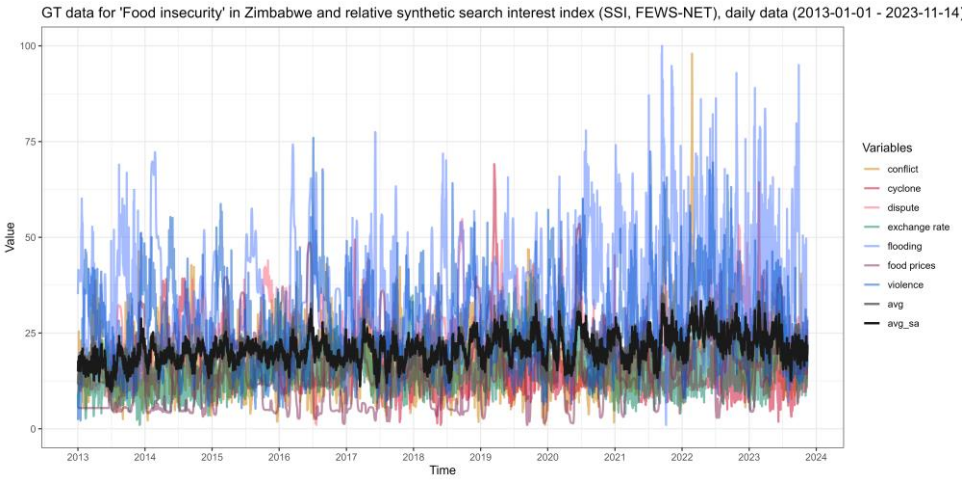


Figure 6. Synthetic search interest index (SSI) for food insecurity in Zimbabwe.

5. Conclusion

The aim of the multi-platform framework for nowcasting is to enable high frequency monitoring of potential food insecurity concerns at the country level, based on machine learning analysis of textual and image data from digitally native sources. Exploratory analysis on data from Zimbabwe point to the feasibility of the models applied to identify food security dimensions in

text and search engine data. Further analysis is needed to interpret converging and diverging trends across the data streams, and their implications to food insecurity early warning.

Acknowledgements

This work was carried out with support from the CGIAR Initiative on Climate Resilience, ClimBeR, and the CGIAR Initiative on Fragility, Conflict, and Migration. We would like to thank all funders who supported this research through their contributions to the CGIAR Trust Fund: https://www.cgiar.org/funders/.

References

Balashankar, A., Subramanian, L., & Samuel P. Fraiberger. (2023). Predicting food crises using news streams. Science Advances, 9(eabm3449). https://doi.org/10.1126/sciadv.abm3449

Carneiro, B., Resce, G., & Sapkota, T. B. (2022). Digital artifacts reveal development and diffusion of climate research. Scientific Reports, 12(14146). https://doi.org/10.1038/s41598-022-17717-8

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. Economic Record, 88, 2–9.

Junnan Li, Dongxu Li, Caiming Xiong, & Steven Hoi. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. ArXiv. https://doi.org/10.48550/arXiv.2201.12086

Jockers, M. L. (2015). Syuzhet: Extract Sentiment and Plot Arcs from Text. https://github.com/mjockers/syuzhet

Queiroz, C., Norström, A. V., Downing, A., Harmáčková, Z. V., De Coning, C., Adams, V., Bakarr, M., Baedeker, T., Chitate, A., & Gaffney, O. (2021). Investment in resilient food systems in the most vulnerable and fragile regions is critical. 2(8), 546–551.

Rogers, R. (2013). Digital Methods. The MIT Press.

Rogers, R. (2015). Digital Methods for Web Research. In R. A. Scott & S. M. Kosslyn, Emerging trends in the social and behavioral sciences: an interdisciplinary, searchable, and linkable resource.

Eichenauer, V. Z., Indergand, R., Martínez, I. Z., & Sax, C. (2022). Obtaining consistent time series from Google Trends. Economic Inquiry, 60(2), 694–705.

WB. (2024). World Development Indicators. Washington, D.C., US. (World Bank. Retrieved from: https://databank.worldbank.org/source/world-development-indicators [Accessed February 1, 2024])

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17823

Unveiling New Insights From Textual Unstructured Big Data in Politics Through Deep Learning

Ufuk Caliskan[1], Angela Pappagallo[2], Francesco Ortame[2], Mauro Bruno[2], Francesco Pugliese[2]

[1] Deutsche Post & DHL, Germany, [2]Italian National Institute of Statistics, Italy.

How to cite: Caliskan, U.; Pappagallo, A.; Ortame, F; Bruno, M.; Pugliese, F. 2024. Unveiling New Insights From Textual Unstructured Big Data in Politics Through Deep Learnin. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17823

Abstract

Over the past decade, social media platforms have undergone significant and rapid expansion. One of the key challenges has been effectively analysing the vast amount of unstructured user-generated data they produce. This research delves into the analysis of Italian Twitter data through the application of advanced deep learning models across three primary objectives: text classification, sentiment analysis, and hate analysis. Five cutting-edge models are evaluated, each utilizing distinct word embeddings. Furthermore, this study investigates the effects of processing emojis and emoticons in Italian tweets on sentiment and hate analysis. We compare model performances and suggest optimized approaches for each task. Finally, we apply these methodologies to real-world Twitter data and present our findings through multiple graphs and statistical analyses. This study demonstrates the possibility of extracting new insights and novel information from unstructured textual Big Data in Politics.

Keywords: *politics, deep learning, artificial intelligence, big data, statistics, sentiment*

1. Introduction

Social media platforms, exemplified by Twitter's massive user base of 330 million monthly users generating 500 million daily tweets, offer rich 'Big Data' replete with opinions and sentiments (Leonowicz-Bukała et al., 2021). Recent advances empower researchers to extract insights, notably in politics, aiding in real-time sentiment analysis. This study delves into Italian tweets, employing deep learning techniques like Word2Vec and Fasttext embeddings, along with models like CNN, LSTM, and RCNN, for political classification, sentiment, and hate analysis. Model performance metrics like accuracy and F1-score are scrutinized, with the best approach applied to test Twitter data, showcased through graphical representations. This

research highlights deep learning's prowess in distilling insights from vast social media datasets (Catanese et al., 2023).

2. Methods

Natural language, essential for human communication, evolves over time without explicit rules like programming languages. Natural Language Processing (NLP), a subset of AI, focuses on enabling computers to understand, manipulate, and interpret human language (Bird et al., 2009). Artificial Neural Networks (ANNs) model biological nervous systems and comprise interconnected neurons across input, hidden, and output layers (Yegnanarayana, 2009). Convolutional Neural Networks (CNNs), initially for computer vision, excel in text classification by extracting features regardless of text position, using convolution and pooling layers (Li et al., 2021). Recurrent Neural Networks (RNNs) maintain state across steps, while Long Short-Term Memory (LSTM) networks address long-term dependency challenges by retaining information through specialized gate mechanisms (Hochreiter and Schmidhuber, 1997). Bidirectional LSTMs (BiLSTM) improve performance by processing input sequences in both directions, accessing forward and backward information at each step (Cui et al., 2018). Attention BiLSTM enhances focus on crucial input elements, aiding in sequence processing (Luo et al., 2018). Recurrent Convolutional Neural Networks (RCNNs) combine advantages of RNNs and CNNs, addressing bias towards later words and window size challenges (Siwei Lai et al., 2015). Word Embeddings represent words as low-dimensional continuous vectors, capturing semantic relationships (Xing et al., 2014). Word2vec, by Google, predicts words from context or context from words, with Skip-gram favored for its robust learning (Mikolov et al., 2013). FastText, an extension of skip-gram, overcomes word2vec limitations by employing character-level embeddings (Bojanowski et al., 2017).

3. Results

This section presents three distinct tasks: Politician Classification, Sentiment Analysis, and Hate Analysis. The objective is to compare the three methods described in the previous sections and construct the optimal model for each task.

3.1 Politics Classifier

The main objective of this task is to create a brief text classifier using official tweets from five prominent Italian politicians. The tweets have been obtained through the Twitter API. To develop the classifier, numerous DL methods have been employed and compared. The best model has been then used to classify tweets from other users, including journalists and newspapers. The main idea of this work is that the language styles and keywords used by politicians can distinguish their texts and display similarities between them and their supporters.

The dataset used in this study includes tweets from the official accounts of five major Italian politicians: "*Giuseppe Conte", "Luigi Di Maio", "Matteo Renzi", "Matteo Salvini"*, and *"Nicola Zingaretti"*. A total of 13,541 tweets have been downloaded. URLs have been removed as they do not provide useful information for this task. Additionally, stopwords and non-alphanumeric characters have been deleted from the Training Set. Single characters and multiple white spaces have been removed, and every character has been converted to lowercase. After the preprocessing steps, the total number of tweets in the candidate Training Set is 8,376. As illustrated in Figure 1, the dataset is heavily imbalanced.

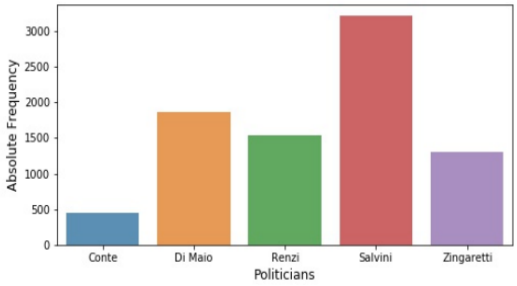


Figure 1. Dataset with Tweets of five major Italian Politicians.

"Giuseppe Conte" has the least Twitter activity with 456 tweets, while *"Matteo Salvini"* has the most with 3214 tweets. Table 1 shows the final performance metrics of the DL models considered by us, including accuracy, precision, recall, and F1-score. Due to the imbalanced condition of the training set, the F1-score is given higher importance.

Table 1. Accuracy, precision, recall and f1-score of all the Deep Learning methods.

W2V-CNN	60,74%	61,76%	60,74%	60,07%
W2V-CNNLSTM	71,12%	70,50%	71,12%	70,63%
W2V-LSTM	**75,48%**	**75,20%**	**75,48%**	**75,19%**
W2V-RCNN	74,64%	74,16%	74,64%	73,77%
W2V-AttBiLSTM	74,58%	74,72%	74,58%	73,94%
FT-CNN	61,99%	61,86%	61,99%	61,34%
FT-CNNLSTM	67,66%	69,33%	67,66%	66,63%
FT-LSTM	64,56%	65,31%	64,56%	63,41%
FT-RCNN	65,10%	65,68%	65,10%	63,06%
FT-AttBiLSTM	54,89%	58,26%	54,89%	51,98%

The table above presents a comparison of data preprocessing (embeddings) and Deep Neural Networks applied to the same Training Set. It is observed that W2V (Word2Vec) outperforms FastText. The CNN (1 Dimensional Convolutional Neural Network) model chosen for this task appears to be unsuitable. The combination CNN-LSTM model performs averagely. However, the LSTM model shows the best performance overall. The W2V-LSTM model has been chosen as the politician classifier, following AttBiLSTM (Bidirectional LSTM with Attention

Mechanisms) and RCNN (Recurrent Convolutional Neural Networks). Then, Official tweets related to Italian newspapers and journalists have been downloaded and classified to measure their closeness to the five politicians previously modelled during training. We have scraped 20,693 official tweets from seven Italian newspapers and 31,538 tweets from eleven Italian journalists. Both datasets have undergone the same preprocessing as the training stage. The W2V-LSTM model has been applied to both datasets in the inference stage. Contingency tables have been calculated and results are projected into a 2-dimensional space using Correspondence Analysis (CA). Canonical analysis (CA) is a multivariate statistical method used to visually represent dependencies in contingency tables. Figure 2-left shows the outcomes of the newspapers, while figure 2-right shows those of the journalists.

3.2. Sentiment Analysis

The objective of the Sentiment Analysis is to create a sentiment classifier for tweets and cross-reference the results with the politician classifier breakthroughs described in the previous section. The chosen Training Set is a combination of three different datasets: Sentipolc data [14], Happy Parents data [47], and Absita data [15]. The Sentipolc data comprises 9,410 Italian tweets, divided into 7,410 tweets for training and 2,000 tweets for the test set.

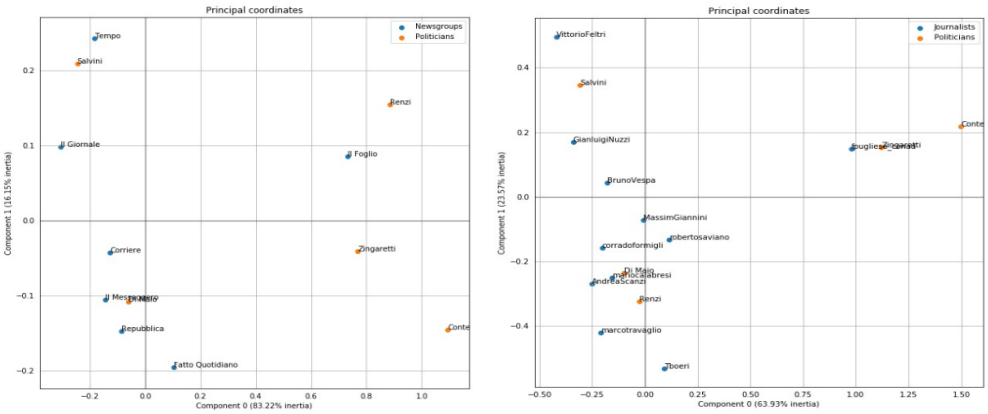


Figure 2. left) Visualization of Correspondence Analysis for Newspapers. The overlapping labels of the points in the bottom left are Il Messaggero for the blue point and Di Maio for the orange point; right) Visualization of Correspondence Analysis for Journalists. The overlapping labels of the points in the top right are fpugliese_conad for the blue point and Zingaretti for the orange point.

The dataset comprises both political and generic tweets, while the test data consists of tweets extracted using hashtags and keywords related to the socio-political topic #labuonascuola. Emoticons and emojis are included in the analysis by replacing them with the corresponding English textual representation using specific libraries (demoji [6]). After training the same deep neural networks as in the previous section, the W2V-RCNN method has been selected as the

best model for this task. We have tested W2V-RCNN on tweets downloaded from politicians to classify their sentiment. Figure 3-left displays the distribution of the sentiment classes for each politician. Subsequently, the sentiment classifier and politics classifier have been jointly applied to tweets containing only the hashtags #primagliitaliani and #lilianasegre. The hashtag #primagliitaliani, meaning 'Italians first', has been actively promoted by Matteo Salvini. Instead, Liliana Segre is an Italian Holocaust survivor and a senator for life in Italy. A discussion about Liliana Segre's protection has been ongoing for years, and the hashtag #lilianasegre has become popular during autumn and winter 2019. We have downloaded 1,086 tweets with the hashtag #primagliitaliani between 18/11/2019 and 27/11/2019, and 1,994 tweets with the hashtag #lilianasegre between 21/11/2019 and 27/11/2019. The final classification of this sub-task consists of two steps. The text describes the distribution of political trends and sentiments in all these tweets. The results are presented in Figure 3-right, which shows that tweets with the hashtag #primagliitaliani are mostly associated with Matteo Salvini, followed by Luigi Di Maio. Classifications of Matteo Renzi and Nicola Zingaretti have been infrequent. Tweets classified as Matteo Salvini are mostly negative or neutral, while tweets classified as Luigi Di Maio are mostly neutral and positive. The tweets that include the hashtag #lilianasegre display a similar pattern, but with less impact from Matteo Salvini on the results.

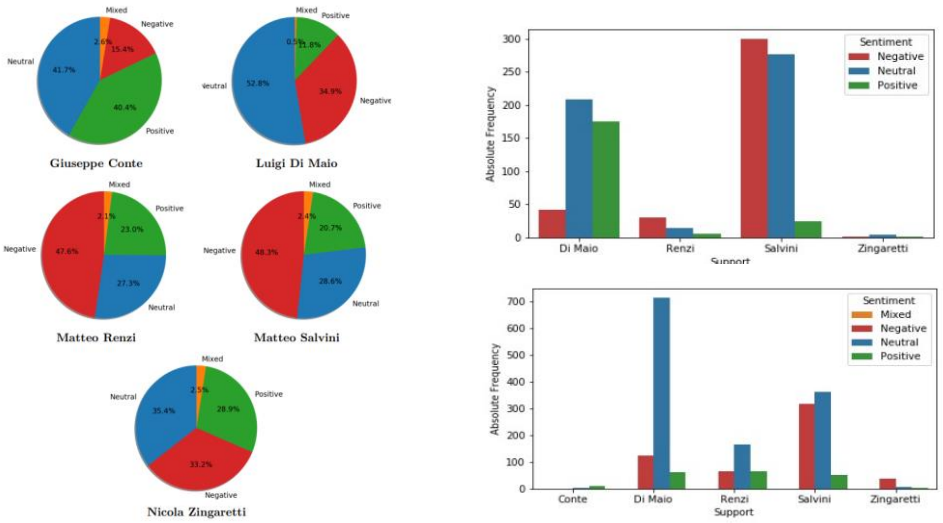


Figure 3.left) Sentiment of the tweets per politician; right) #primagliitaliani (up) and #lilianasegre (down) sentiment and politics simultaneous classification

3.3. Hate Analysis

The final task is building an Italian hate classifier, addressing the surge of online hate expression, notably on social media. Limited labeled datasets exist for hate classification,

mirroring sentiment analysis challenges. Using Facebook comments and Twitter tweets, the study aims to analyze and apply findings to political figures. The hate recognition data utilized in this study originates from the EVALITA 2018 Hate Speech Detection Task, encompassing Facebook and Twitter datasets. For this datase, Facebook comments have been collected from public pages of Italian newspapers, politicians, artists, and groups, suspected to contain hateful content. A total of 99 posts have generated 17,576 comments, annotated by five students into classes: no hate, weak hate, and strong hate. Annotations have been were simplified for the EVALITA 2018 task, resulting in two classes: 0 for no hate and 1 for hate, with 4,000 posts in total. As with sentiment analysis, the preprocessing applied to the hate data has the steps, the emoticons translation included. After the training, the best model has been the W2V-RCNN, as in sentiment analysis. The hate classifier has been applied to tweets directed at the five politicians. For each politician, 2000 tweets with a minimum number words equal to three have been downloaded, in which they have been mentioned. The data have been collected between 25/11/2019 and 27/11/2019. After preprocessing, 5,937 tweets have been removed from the corpus of 10,000 tweets. By exploiting the trained hate classifier, each tweet from the remaining corpus of 4,063 newly download tweets has been classified as hateful or not hateful. In Figure 4 we depict the hate ratio of each politician during the reference period, which is calculated by dividing the number of hateful tweets by the total number of tweets.

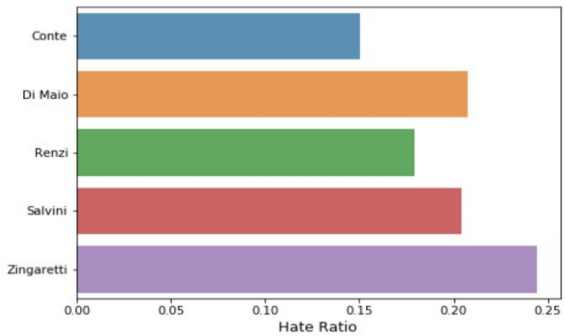


Figure 4 Distribution of the tweets with mentions

The hate classifier has identified that Nicola Zingaretti received the highest percentage of hateful tweets at 25%, followed by Matteo Salvini and Luigi Di Maio at 20%. Meanwhile, Matteo Renzi received 18% and Giuseppe Conte received 15% of hateful tweets in which they were mentioned. Finally, a graph was plotted to display the relationships between various politicians based on the negative sentiment expressed in tweets mentioning one politician by another, as shown in Figure 5. The thickness of the arrow indicates the strength of the negative relationship between two politicians or their supporters. This suggests that the thickness of the arrow may indicate a deterioration in the relationship between two politicians, while a slimmer arrow may indicate an improvement in their relationship.

4. Conclusions

This research conducts a comparative analysis of various methodologies to determine the most effective Deep Learning (DL) approach for classifying political tweets, conducting sentiment

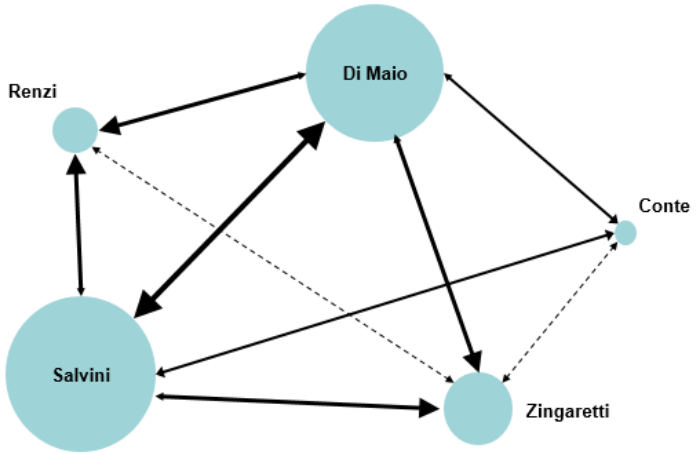


Figure 5 Distribution of the tweets with mentions

analysis, and detecting hate speech within tweets. Each task uses a different DL method specifically designed for tweet analysis. The W2V-LSTM method has been shown to perform well in classifying political tweets, achieving an F1-score of 75.19% on a dataset of tweets from five prominent Italian politicians. The study indicates that tweets from various users, especially shorter ones, demonstrate a writing style that is more closely associated with a particular political class, which affects classifier decisions. To improve classifier evaluation in future studies, it is recommended to establish a ground truth using tweets from users who express political opinions. Additionally, expanding the dataset to include more tweets and politicians is advised. The W2V-RCNN method outperforms others in sentiment analysis, achieving an F1-score of 77.58%. However, the inclusion of various datasets may skew findings, highlighting the need for a larger, labeled corpus for accurate evaluation. Additionally, further research is warranted to enhance analysis precision, particularly in processing emojis and emoticons and their wider range and equivalent translations. This work demonstrates the use of advanced Artificial Intelligence tools, specifically Deep Neural Networks (Deep Learning), to extract valuable insights from unstructured textual data such as tweets and short texts. These models provide significant benefits in modern data analysis as they can assist politicians (or individuals in general) in extracting additional information from textual data that would otherwise be unattainable from other sources. This creates innovative opportunities for data analysis in both Official Statistics and the analysis of the orientations of a reference population. In the future, more recent analysis models such as Transformers (Vanilla Transformer, Bert, GPT) could be

used. These models may achieve superior prediction metrics and provide better analysis of relationships between politicians with more sophisticated hate and sentiment analysis.

References

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the association for computational linguistics, 5, 135-146.

Catanese, E., Bruno, M., Stefanelli, L., & Pugliese, F. (2023). Measuring Social Mood on Economy during Covid times: effects of retraining Supervised Deep Neural Networks.Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.".

Cui, Z., Ke, R., Pu, Z., & Wang, Y. (2018). Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction. arXiv preprint arXiv:1801.02143.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

Leonowicz-Bukała, I., Adamski, A., & Jupowicz-Ginalska, A. (2021). Twitter in Marketing Practice of the Religious Media. An Empirical Study on Catholic Weeklies in Poland. Religions, 12(6), 421.

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems, 33(12), 6999-7019.

Luo, L., Yang, Z., Yang, P., Zhang, Y., Wang, L., Lin, H., & Wang, J. (2018). An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. Bioinformatics, 34(8), 1381-1388.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Ming, Y., Cao, S., Zhang, R., Li, Z., Chen, Y., Song, Y., & Qu, H. (2017, October). Understanding hidden memories of recurrent neural networks. In 2017 IEEE conference on visual analytics science and technology (VAST) (pp. 13-24). IEEE.

Raffel, C., & Ellis, D. P. (2015). Feed-forward networks with attention can solve some long-term memory problems. arXiv preprint arXiv:1512.08756.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In Twenty-ninth AAAI conference on artificial intelligence, 2015.

Xing, C., Wang, D., Zhang, X., & Liu, C. (2014, December). Document classification with distributions of word vectors. In Signal and information processing association annual summit and conference (APSIPA), 2014 asia-pacific (pp. 1-5). IEEE.

Yegnanarayana, B. (2009). Artificial neural networks. PHI Learning Pvt. Ltd..

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17824

The use of non-official data source for the analysis of public events: evidences from the Eurovision Song Contest 2022

Alessia Forciniti[1], Andrea Marletta[2], Magda Moretti[2]

[1]Department of Humanities, IULM University, Milan, Italy, [2]Department of Economics, Management and Statistics, University of Milano-Bicocca, Milan, Italy.

How to cite: Forciniti, A.; Marletta, A.; Moretti, M. 2024. Paper The use of non-official data source for the analysis of public events: evidences from the Eurovision Song Contest 2022. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17824

Abstract

The use of non-official data sources as Twitter has been implemented for the monitoring of social and public events in many different fields during last years. Following this issue, this work proposes to analyse a very well-known musical event, the Eurovision Song Contest (ESC) 2022 using tweets pooled by the official hashtag of the competition. From a methodological point of view, text mining techniques have been applied to detect the most influencing terms and topics tweeted by users during the show and to compare the official results of the contest with a ranking only based on the appreciation of the Twitter users on posts relative to the participant countries.

Keywords: *Twitter Data; Eurovision Song Contest; Text Mining.*

1. Introduction

The statistical analysis of public events has always been topic of interest of quantitative stakeholders from many points of view. If for these events a competition is considered, the study of the factors influencing the final result covered many contributions. For example in sport statistics, the best predictions are achieved taking into account a very huge quantity of statistics from official sources.

On the other hand, during last years, the social media have been recognized as one of the most powerful data source able to add a substantial contribution in forecasting procedures. The decision to place a like or to share a content on social media can be interpreted not only as a form of appreciation but also like the expression of a behaviour. Starting from this, the stakeholders can use this information in the decision making process. It is important to note that social media data are not necessarily quantitative, as number of likes, shares and views but they can also involve qualitative aspects as comments, opinions and answer to open questions making the single word a vector of information.

The contact point between the analysis of public events and the rising communicative power of the social media, lies in the possibility to arrange a link between the event of interest and the social media users using a common hashtag. In this way, all the people are following the public event live or online could comment and/or express their appreciation or dislike in real time. For some events as Tv show or musical events in which a competition is provided, this new form of interaction is translated into a voting system that could influence the final result of the competition (Demergis, 2019, Kumpulainen et al., 2020).

The aim of this study is to provide an analysis based on Twitter data of the Eurovision Song Contest (ESC) 2022 using the tweets collected by the official hashtag of the competition in order to detect the most influencing topic and terms used by Twitter users during the show. The ESC is a very well-known musical event organized yearly by the European Broadcasting Union (EBU), it is a competition among countries in which each country is represented by an artist and the winner country is declared after a voting system based on a combined ranking using quality juries and phone-in-vote. Another objective of this study is to compare the final result of the contest with an alternative ranking only based on the tweets in which the participant country is mentioned. The final goal is to understand if a correspondence exists between the real and a Twitter-based ranking in order to measure the different perception of the Twitter population respect to the real world for this event.

The paper is structured as follows: after the introduction, a second section is dedicated to the methodologies used to answer the research objectives. A third section will show the description of the dataset and some preliminary results. Finally, some conclusions will follow.

2. Text Mining techniques

The authomatic ananlysis of textual data, also known as text mining, is a set of tools allowing the traslation from word to information. This technique is based on the availability of textual and non-structured data, that after some operations of data cleaning can be elaborated to be transformed in structured data, able to be necessarily analysed (Tuzzi, 2003).

A complete definition of text mining can be retrieved in Feldman and Sanger (2007): 'Text mining is a new and exciting area of computer science research that tries to solve the crisis of information overload by combining techniques from data mining, machine learning, natural language processing, information retrieval and knowledge management'. In particular, when the text mining objects are referred to social networks, this technique is also known as web mining or social media mining.

The starting point of the authomatic ananlysis of texts is the corpus, that is, the set of textual units coherent to be analysed (Bolasco and De Mauro, 2013). It is therefore a homogeneous collection for treated topic, structure of the sentence and length able to be observed without

systematic distortions. On the basis of the extent and of the composition of the corpus, two typologies of corpus are identified: collection of texts and collection of fragmented texts. A second classification of the corpora detected a single document or a set of records in which each row is differently generated. For social media text mining the corpus is a collection fo fragmented text with a set of rows, that are posts, tweets, reviews generated by social media users.

Before starting the textual analysis, a pre-procssing procedure is needed using some enhancements useful for this category of data. Firstly, a process of micro-fragmentation (tokenization) of the corpus is applied linking a numerical code to each word. Secondly, all the uppercase characters are detected and transformed into lowercase, this step simplifies the analysis but it could generate a loss of information. Thirdly, all the numerical string and the separators (comma, dot, etc..) has to be deleted. Fourthly, some entire words could be deleted from the corpus, they are empty enitities and their elimination did not create loss of meaning. This list usually contains articles, conjunctions, prepositions prononuns, possessive adjectives, common verbs. Finally, the Term-Document Matrix (TDM) is composed by all the words that are still in the corpus after the application of the previous steps. In the TDM, each row is a word, each column is a document of the corpus and each cell is the occurrency of the single word in the single document.

After the pre-processing phase, in order to extract information from textual data, the existing techniques can be classified into three groups: the frequency analysis, the cluster analysis and the sentiment analysis. The frequency analysis consists in the counting of the words contained in the corpus in order to create a dictionary of most important terms. This kind of analysis is descriptive and not sufficient to catch the total information from the corpus. Beyond the removed stopwords, the terms with the highest frequency are often generic and expected because related to the interest topic. For this reason, in this study the attention is focused only on some categories of words as countries and artists. The cluster analysis regards the possibility to create groups of terms that recur together using hierarchical and non-hierarchical models already used for quantitative data. The Sentitment Analysis (SA) represents a discipline based on Natural Language Processing (NLP) and aimed at identifying opinions, emotional, behavioral and attitudinal dimensions expressed in natural language texts (Pang and Lee, 2008). The main objective is to determine the semantic orientation of texts written in natural language by classifying documents based on their polarity; where polarity refers to the linguistic distinction between affirmative and negative terms. There are three different levels of semantic ordering characterized by different granularity: a) subjectivity/objectivity (SO orientation): aimed at determining factual nature or subjective judgment; b) positivity/negativity (PN orientation) (Liu, Hu & Cheng, 2005): with the purpose of determining whether it expresses a positive or negative opinion, and often neutral; c) strength of positivity/negativity (PN strength):

aimed at indicating different levels of intensity of positivity and negativity; from the most negative to the most positive (Nielsen, 2011).

The three different levels of semantic analysis can be performed on the individual document, at the sentence level, and at the aspect level. In the first case, the positive or negative polarity is returned at the general document level, while in the second case, the document is segmented into sentences and the semantic orientation is detected for each sentence. Polarity attribution can occur through lexicon-based approaches (generated manually or automatically, semi-automatically), machine learning, deep learning, hybrid strategies, or those recurring to artificial intelligence. The choice of the lexicon to adopt represents an ongoing debate, and the application context is also an area of continuous research, dividing lexicons into generalists and domain-based. Further discussions in the literature focus on language, as most lexicons are developed for English, not covering other languages or providing limited resources (Zavarrone & Forciniti, 2023). Another possibility is represented by using multilingual language models, that classify and cluster tweets into homogeneous groups irrespective of their language, thereby facilitating a more comprehensive summarization of opinions expressed across various languages.

In this study, a strategy of lexicon-based classification at the document level is adopted, where each tweet is a different document. Specifically, the NRC (National Research Council of Canada) lexicon developed by Mohammad and Turney (2010; 2013) for sentiment analysis of Twitter, and available for various languages including Italian, was used. The NRC lexicon allowed the detection of PN orientation: positive, negative, and neutral opinions were detected as -1, 0, +1 to indicate respectively negative, neutral, and positive terms. Additionally, we utilized the word–emotion association developed by Mohammad and Turney through Amazon's Mechanical Turk. The NRC is based on annotations of the eight primary emotions suggested by Plutchik (1962, 1994): joy, sadness, anger, fear, disgust, surprise, trust, and anticipation.

3. Data collection and results

Data have been collected through the official Twitter API using all the tweets containing the official Italian hashtag of the competition #ESCita. The event took place in Turin, Italy from 10th to 14th May 2022 in three shows. The contest was won by Ukraine Since the event has a duration of more days with two semi-finals show and a final show, multiple extractions of tweets have been achieved starting from the day of the first semi-final until to the day after the final. For this study, only tweets about the day of the final show have been considered for a total of n=61,464 tweets. The variable of interests for each single tweet are: the text in Italian language, the exact time of publication, the username of the author, the number of likes and the number of retweets.

As reported in the previous section, a huge procedure of data cleaning has been applied on the text variable before obtaining the TDM, after the transformation into lowercase and the removal

of numerical string and the separators (comma, dot, etc..), a list of words has been added to the usual stopwords list, because marked as empty terms. For example all the terms directly related to the hashtag #ESCita like "eurovision, esc, eurovsiontv, eurovisionsongcontest, esc, eurovisionrai" have been removed as the stopwords. After the construction of the TDM, a dictionary of the entire corpus composed by 33,107 words has been obtained. The word with the maximum frequency in the dictionary is "canzone" (4,403 times) the Italian term standing for "song", this could be expected since the ESC is primarly a music contest. In the top-10 of the most present words, there are also three countries (Spain, Ukraine and Italy), two Italian presenters (Malgioglio and Laura) and two Italian singers (Blanco and Mahmood). As expected, the dictionary is very focused on the Italian side of the competition, this is a limitation of the study but since the show take place in Italy, this choice could be justified.

A good way to synthetize the entire dictionary is the use of a wordcloud, a graph in which all the words over a fixed threshold are represented with a different colour and size, the terms with a bigger type are those with a bigger frequency. In figure 1, all the terms with at least 600 presences in the entire corpus have been represents using a wordcloud. Looking at this figure, it is possible to note that most of the terms indicate name of the participant countries or artists involved in the show as presenters, singers and guests.

Figure 1. Wordcloud of the terms with at least 600 presences in the ESC textual corpus. Source: Twitter Eurovision data (2022).

This issue advices to focus the attention on some specific categories of words, for example the participant countries, in order to search for a link between the final ranking of the contest and the appreciation that these countries received from Twitter users that used the #ESCita hashtag.

The final ranking of ESC has been obtained by the sum of points of televoters and juries from all the participant countries, with the exception that each country can not to vote for itself, therefore the Italian voters ranking did not consider Italy. A simplified version of the obtained dictionary only considering the mentioned countries could be considered as a proxy of the appreciation ranking by Twitter users. The produced ranking has many distortions because it only considers tweets using an Italian hashtag and above all because a mention in a post do not correspond necessarily to an appreciation. Nevertheless all these presuppositions, a simple correlation coefficient between the official ranking and the mentions ranking has been computed and it is equal to 0.52 (p-value = 0.007) denoting a positive association between the two rankings. In Figure 2, a simple scatterplot between the two rankings denotes some potential clusters of country, for example France and Germany are in bottom right part of the scatterplot because nevertheless a very bad position in the official ranking, they have a high number of mentions in Italian tweets, probably because they are in the big 5 group (a group of countries directly admitted into the final show) and a lot of users commented their disappointing result. Other correlations between rankings have been computed, for example comparing the mention ranking with the televoters ranking in Italy (available on the official website of ESC: https://eurovision.tv/), to avoid the bias due to the fact that the mention ranking is only based on Italian tweets, but the correlation coefficient is still about 0.50 (p-value = 0.012).

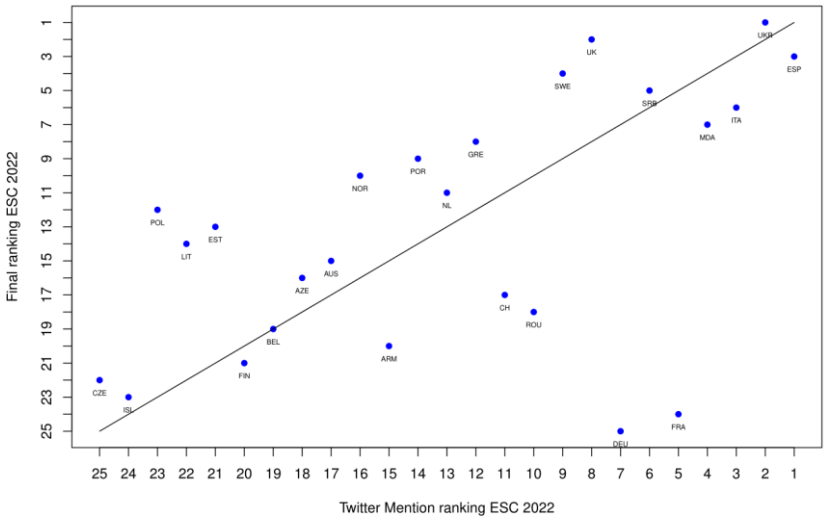


Figure 2. Comparison between official ranking ESC and a ranking based on Twitter mentions. Source: Twitter Eurovision data (2022).

Finally, a sentiment analysis on the entire corpus of the final show of ESC has been realized using the NRC lexicon, according to this classification each tweet could be categorized as positive, negative or neutral. The 91.3% of the tweets are classified as neutral tweets, the 5.1% as positive tweets and the remaining 3.6% as negative tweets.

In figure 3 a representation of the sentiment analysis is about the eight primary emotions described in the previous section. The anticipation is most present emotion, it is present in 20.1% of the tweets, the trust is following with the 18.9%, followed by sadness and joy with respectively 12.9% and 12.5%. A limitation of this study is represented by the fact that sentiment analysis doesn't consider any irony or sarcasm detection: irony is really important when managing comments about TV contests.

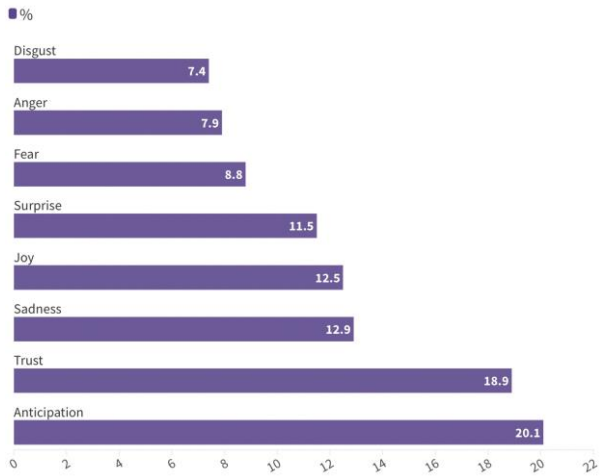


Figure 3. Eight primary emotions for tweets about the final show of ESC. Source: Twitter Eurovision data (2022).

4. Conclusions

The aim of this work was to analyse a very well-known musical event as the ESC 2022 using non-structured data from social media. In particular, a text mining analysis has been conducted on 61,464 tweets containing the official hashtag for Italy, #ESCita and posted during the last day of the event, detecting some preliminary evidences. From a methodological point of view, after an important pre-processing procedure of data cleaning, a dictionary has been created identifying the participant countries as relevant category of words. On the basis of the mentions received by the countries, a proxy of an appreciation ranking has been compared with the official ranking denoting a good correlation. Finally a sentiment analysis on the entire corpus revealed the anticipation as main emotion in the tweets. The study is still in progress and many limitations are present, both for the nature of data and because a direct connection between mentioning and appreciation is missing. Following this issue, some future works could regard the enhancement of the mentions ranking taking into account like and retweets received by mentioned posts and introducing the publication hour as control variable.

References

Bolasco, S., & De Mauro, T. (2013). *L'analisi automatica dei testi: fare ricerca con il text mining*. Carocci editore.

Demergis, D. (2019, October). Predicting Eurovision song contest results by interpreting the tweets of Eurovision fans. In 2019 Sixth international conference on social networks analysis, management and security (SNAMS) (pp. 521-528). IEEE.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press.

Kumpulainen, I., Praks, E., Korhonen, T., Ni, A., Rissanen, V., & Vankka, J. (2020, September). Predicting Eurovision Song Contest Results Using Sentiment Analysis. In Conference on Artificial Intelligence and Natural Language (pp. 87-108). Cham: Springer International Publishing.

Liu, B., Hu, M., & Cheng, J. (2005). Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*.

Mohammad, S., & Turney, P. (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, *29*(3), 436-465.

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, *2*(1–2), 1-135.

Plutchik, R. (1962). *The Emotions*. New York: Random House.

Plutchik, R. (1994). *The psychology and biology of emotion*. HarperCollins College Publishers.

Tuzzi, A. (2003). *L'analisi del contenuto. Introduzione ai metodi e alle tecniche di ricerca*. Carocci editore.

Zavarrone, E., & Forciniti, A. (2023, July). CSR & Sentiment Analysis: A New Customized Dictionary. In *International Conference on Deep Learning Theory and Applications* (pp. 466-479). Cham: Springer Nature Switzerland.

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17782

A Bibliometric Study of Stakeholder Opinion Mining and Sentiment Analysis in Crisis Communication

Homa Molavi, Lihong Zhang
Department of Engineering Management, School of Engineering, The University of Manchester, UK.

How to cite: Molavi, H.; Zhang, L. 2024. A Bibliometric Study of Stakeholder Opinion Mining and Sentiment Analysis in Crisis Communication. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17782

Abstract

In the contemporary landscape, the ability to effectively manage crises and communicate with stakeholders is paramount for organizations. As the frequency and complexity of crises continue to escalate, understanding stakeholder opinions and sentiments becomes increasingly crucial for crafting timely and appropriate responses. This bibliometric study delves into the landscape of stakeholder opinion mining and sentiment analysis within crisis communication, aiming to discern trends, identify key contributors, and uncover potential gaps in the existing literature. Leveraging data from the Scopus database from 2004 to 2024, the analysis reveals a notable increase in publications over time, particularly since 2019, highlighting the growing interest in this field. The United States, the United Kingdom, and Germany emerge as leading contributors, while institutions such as The University of Texas at Austin and Universiteit van Amsterdam demonstrate significant productivity. However, limited collaboration between top institutions and authors suggests opportunities for enhanced knowledge exchange and interdisciplinary collaboration.

Keywords: *Crisis Communication; Bibliometric Mapping; Opinion Mining; Sentiment Analysis; Bibliometric Review; Stakeholder Behaviour.*

1. Introduction

Effective communication is a cornerstone of organisational success, permeating interactions among its members, regardless of their formal or informal nature (Rahimnia & Molavi, 2021). However, in today's dynamic digital landscape, organisational communication faces unprecedented challenges during crises, jeopardizing stakeholder relationships and operational continuity. Stakeholders, wielding significant influence, can either bolster organisational endeavors or pose substantial challenges (Degtjarjova et al., 2018). The integration of stakeholder theory with strategic thinking has long been pivotal in organising critical

information for strategic planning, enhancing the efficacy of business policies and strategies (Freeman et al., 2020).

During crises, stakeholder behaviour profoundly shapes organisational outcomes. Research underscores that stakeholders' actions are often driven by their perceptions of an organisation's response, disseminated through intricate feedback processes involving multiple stakeholders (Bosse et al., 2009; Larson, 1992). Thus, understanding stakeholder opinions and sentiments is paramount for crafting timely and effective responses. The accuracy of organisational assumptions about stakeholder behavior during crises determines the ability to avert or mitigate their impact (Alpaslan et al., 2009; Mitroff & Kilmann, 1984; Nathan & Mitroff, 1991; Pearson & Clair, 1998; Perrow, 1999; Ulmer, 2001).

Stakeholder behavior, encompassing both cooperative engagement and potential threats, offers insights into their impact on crisis response (Savage et al., 1991). This understanding empowers decision-makers to develop strategies for managing stakeholders and safeguarding organization (Mwesigwa et al., 2018; Nguyen & Rose, 2009). Furthermore, the proliferation of social media platforms and online forums has amplified stakeholders' voices, significantly influencing public perception and organisational outcomes during crises.

This study employs bibliometric analysis to discern trends in stakeholder opinion mining and sentiment analysis within crisis communication. Through this method, the research endeavors to discern pivotal trends, and detect gaps. The ultimate goal is to enrich our understanding of the evolutionary trajectory and current state of knowledge concerning stakeholder opinion mining and sentiment analysis during crises.

To conduct a comprehensive bibliometric analysis on the corpus of literature pertaining to stakeholder opinion mining and sentiment analysis within crisis communication, the following questions will be addressed:

1. **RQ1: Trend Analysis**: What are the emerging keywords and prevailing trends in the corpus of literature concerning stakeholder opinion mining and sentiment analysis in crisis communication?
2. **RQ2: Global Leadership and Collaborative Dynamics:** Which countries, scholars, and institutions are at the forefront of research on stakeholder opinion mining and sentiment analysis in crisis communication? and how do their collaborative networks contribute to the overall knowledge development?
3. **RQ3: Scholarly Impact and Collaboration Networks:** What are the prominent co-citation patterns and bibliographic coupling connections among these influential works?

2. Methodology

Bibliometric analysis, described as a collection of mathematical and statistical techniques, serves as a valuable tool to showcase the latest developments and ongoing insights within a particular research area. This method provides an efficient means to uncover the underlying intellectual framework of a research field or subject matter, facilitating a deeper understanding of its structure and trends (Molavi & Zhang, 2024).

2.1. Keywords Search

Utilising the Scopus database, our search strategy using keywords "(crisis OR covid OR pandemic) AND (opinion OR sentiment OR text) AND (communicat*)" yielded a total of 5062 documents. To curate the final collection of highly cited articles, a set of inclusion and exclusion criteria guided the selection process. Specifically, articles published between 2004 and 2024 were considered for inclusion. Additionally, the discipline of publications was restricted to areas such as engineering, decision sciences, multidisciplinary studies, accounting, social science, business, or management. Furthermore, only sources written in English were included.

After applying these criteria, a total of 1029 documents were obtained for the next step in the analysis.

3. Results

This section presents the results of the analyses utilising VOSviewer software.

3.1. Analysis of Trends in the Corpus of Literature

Fig 1 illustrates the number of papers published in the field of stakeholder opinion mining and sentiment analysis within the realm of crisis communication from 2004 to 2024. The trend analysis reveals an upward trend in the number of documents published per year. Notably, the data for the year 2024 is partially considered as the research was conducted in March 2024. Since 2019, there has been a significant increase in the number of publications, rising from 28 documents to 222 articles in 2023. Even within the first three months of 2024, 44 publications have been recorded, nearly double compared to 2015. This trend underscores the growing interest and relevance of stakeholder opinion mining and sentiment analysis in the context of crisis communication.

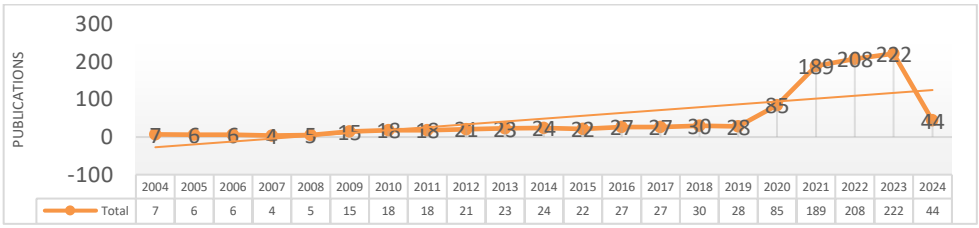


Fig 1. Trends in annual paper volume

3.2. Country-Specific Analysis

The country-specific analysis highlights the top ten countries or regions with the highest number of publications, as shown in Table 2. The United States leads with 266 documents and 5555 citations, followed by the United Kingdom with 92 documents and 1434 citations. Germany ranks third with 59 documents and 1027 citations. These findings underscore the significant contribution of these countries to documents within the analyzed dataset (Fig 2).

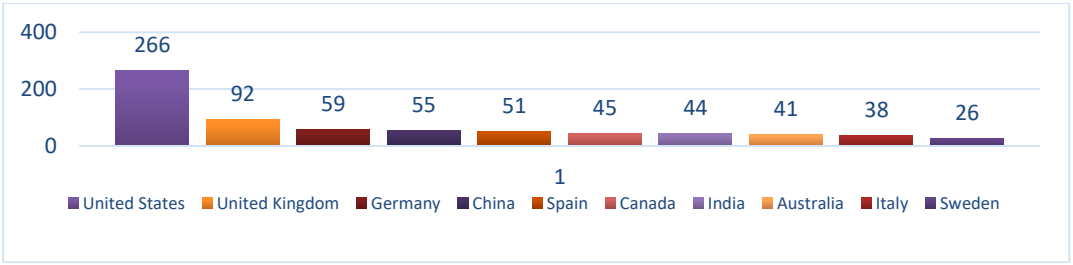


Fig 2. Top 10 article-producing countries.

3.3. Analysis of Institution's Contribution

The analysis of institutional contribution highlights The University of Texas at Austin, Universiteit van Amsterdam, and University College London as the most productive institutions. Despite their notable individual contributions, there were no partnerships or links identified between these institutions. This suggests that while these institutions are independently prolific in their research output, there is minimal or no collaboration between them.

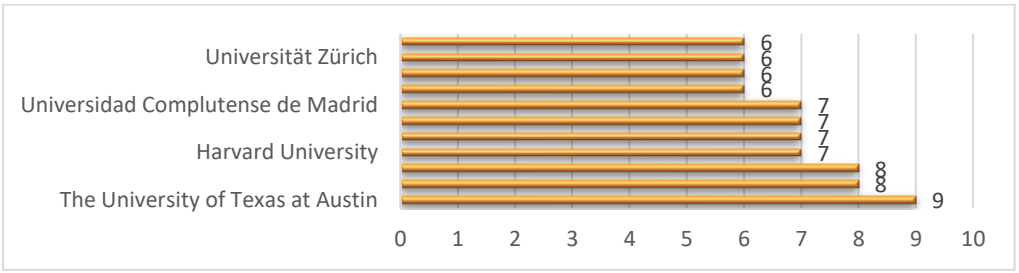


Fig 3. Top 10 Institutions' Contributions

3.4. Analysis of Leading Scholars and Collaborative Networks

Out of 2,842 authors analysed, only six met the threshold of having at least three documents each. These authors include Eisele, O. (5 documents), Arce-García, S. (3 documents), Massarani, L. (3 documents), Stieglitz, S. (3 documents), and Weber, F. (3 documents). However, there was no evidence of collaboration between these authors.

3.5. Keyword Analysis

The keyword analysis reveals the top 15 high-frequency keywords based on a minimum occurrence threshold of 40. Covid-19" dominates with 280 mentions, followed by "social media" with 162. "Pandemic" is significant with 121 occurrences, while "crisis communication" and "coronavirus disease 2019" each appear 81 times. "Communication" is mentioned 80 times, and "public opinion" follows closely with 71 instances. Other notable terms include "interpersonal communication" (67), "epidemiology" (59), and "twitter" (45). "Public health" and "sars-cov-2" round off the list with 44 and 40 mentions, respectively.

This analysis provides insight into the most frequently mentioned keywords, highlighting the dominant themes and topics within the dataset. The emphasis on terms related to COVID-19, social media, crisis communication, and public health underscores the significant focus on these subjects.

From the extracted keywords, several conclusions and insights can be drawn regarding the prevailing themes and topics within the dataset:

1. Dominance of COVID-19: The high frequency of keywords such as "covid-19," "coronavirus disease 2019," and "sars-cov-2" indicates a strong emphasis on the COVID-19 pandemic.

2. Impact of Social Media: The prominence of "social media" as a keyword highlights the significant role of social networking platforms in shaping discourse, communication, and dissemination of information during the pandemic. The inclusion of keywords like "interpersonal communication" and "twitter" suggests a focus on interpersonal interactions and social networking platforms in the context of the pandemic. This indicates potential research or

discussions on how individuals communicate, exchange information, and engage with others during times of crisis.

3.6. Citations

As illustrated in Fig 4, the article by Norris in 2011 stands out as the most cited, amassing an impressive 1,329 citations. Following closely is Ferri's work in 2020, which has garnered 446 citations, and Kavanaugh's article from 2012 holds a substantial position with 414 citations. These citation counts highlight the significant impact and scholarly recognition that these particular articles have received within their respective fields, emphasizing their contributions to the academic discourse.

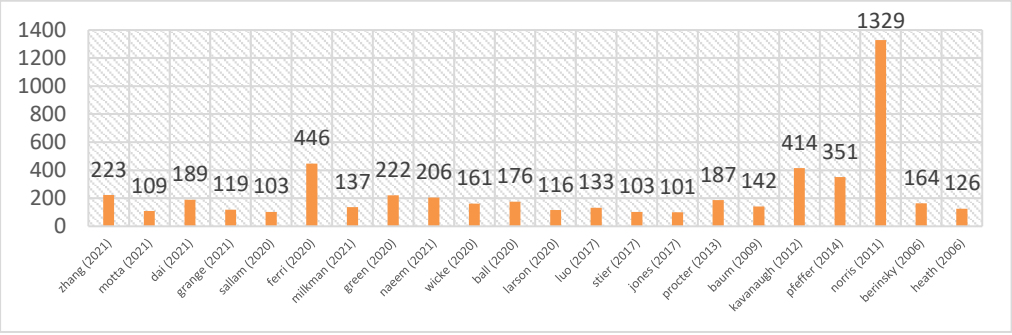


Fig4. Citations

3.7. Co-citation

Out of the 48,001 cited references, only 11 surpass the minimum citation threshold of 13. Notably, the cited references that meet or exceed this criterion vary in citation counts and total link strength. One example is the work of Benoit W.L. titled "Image Repair Discourse and Crisis Communication" published in Public Relations Review in 1997, which garnered 16 citations and a total link strength of 16. On the contrary, Braun and Clarke's "Using Thematic Analysis in Psychology" from Qualitative Research in Psychology in 2006, with 13 citations, shows a minimal total link strength of 0 (Fig5). The diversity in citation counts and link strengths among these references underscores the nuanced impact and interconnectedness within the scholarly network surrounding these specific works.

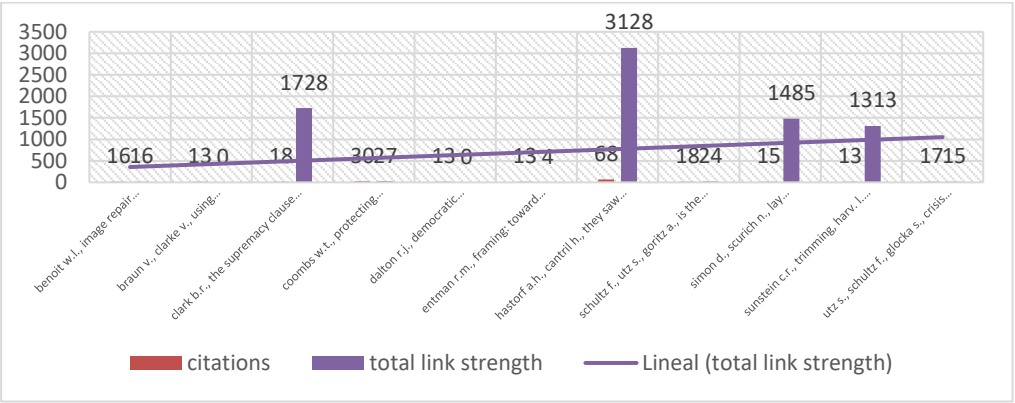
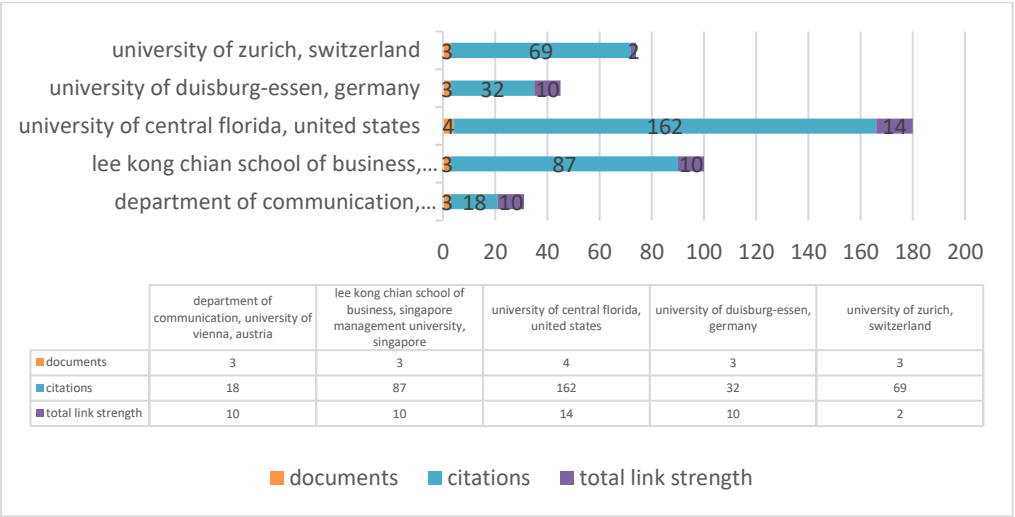


Fig 5.Co-citation

3.8. Bibliographic Coupling

The Fig 6 summarises the bibliographic coupling metrics for each organization. With four documents and 162 citations, the University of Central Florida displays a considerable level of bibliographic coupling. The high citation count suggests how many times documents from each organization cited the same third-party documents. Additionally, its total link strength of 14 indicates a strong interconnectedness among the cited documents, reflecting a cohesive research network within the institution. By examining the collaborative dynamics revealed in the analysis, researchers can identify potential partners for collaboration. Institutions with similar research interests and high bibliographic coupling metrics may benefit from collaborative efforts to further advance research in the field.



	department of communication, university of vienna, austria	lee kong chian school of business, singapore management university, singapore	university of central florida, united states	university of duisburg-essen, germany	university of zurich, switzerland
documents	3	3	4	3	3
citations	18	87	162	32	69
total link strength	10	10	14	10	2

Fig 6. Bibliographic Coupling

3.9. Documents by Subject Area

The findings of the study unveiled a diverse distribution of documents across various disciplines. Among them, Social Sciences dominated with 715 documents, constituting 59.5% of the total. Following closely behind was 'Business, Management, and Accounting' with 222 documents, showcasing a substantial presence. Additionally, Multidisciplinary fields contributed 129 documents, while Engineering and Decision Sciences accounted for 113 and 23 documents respectively, demonstrating a varied landscape of academic inquiry. These statistics underscore the interdisciplinary nature of research and highlight the multifaceted interests within scholarly pursuits (Fig 7).

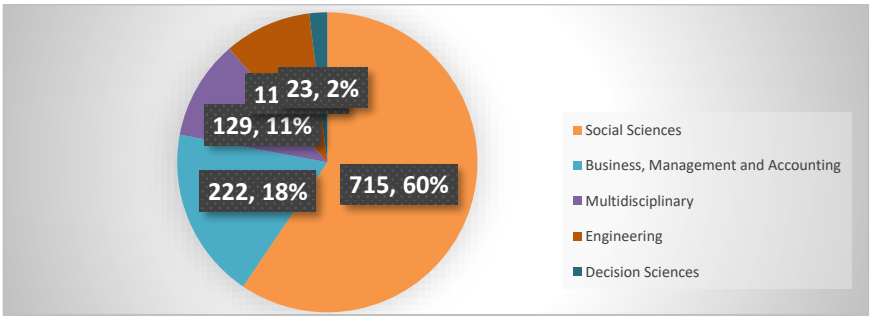


Fig 7. Documents by Subject Area

4. Discussion

RQ1: Trend Analysis

The findings of this bibliometric analysis shed light on several significant aspects of stakeholder opinion mining and sentiment analysis in the context of crisis communication. The trend analysis reveals a consistent increase in publications from 2015 to 2023, with a significant surge in 2024. This trend indicates a growing interest in stakeholder opinion mining and sentiment analysis within crisis communication, reflecting the evolving landscape of organisational communication strategies in response to crises.

The prominence of keywords related to COVID-19, social media, and crisis communication reflects current research priorities. However, the relatively low frequency of keywords related to specific stakeholder groups suggests a potential gap in understanding nuanced stakeholder dynamics within different contexts.

RQ2: Global Leadership and Collaborative Dynamics

The dominance of the United States, the United Kingdom, and Germany in publications underscores their leadership role in advancing research in this field. However, the lack of

substantial contributions from other countries suggests a potential gap in global collaboration and knowledge dissemination. While institutions like The University of Texas at Austin and Universiteit van Amsterdam demonstrate significant productivity, the absence of collaboration between top institutions indicates a missed opportunity for synergistic research efforts. Closing this gap could lead to more comprehensive and impactful insights into stakeholder opinion mining and sentiment analysis. The identification of prolific authors offers valuable insights into individual contributions, but the absence of collaborative networks among them highlights a potential gap in shared research endeavors. Fostered collaboration could lead to more innovative and comprehensive approaches to understanding stakeholder behavior during crises.

RQ3: Scholarly Impact and Collaboration Networks

While certain articles receive significant citation counts, the lack of strong co-citation patterns among seminal works may indicate a fragmented research landscape. The identification of prolific authors offers valuable insights into individual contributions, but the absence of collaborative networks among them highlights a potential gap in shared research endeavors. The presence of bibliographic coupling indicates a cohesive research network within certain institutions, such as the University of Central Florida. However, the absence of widespread coupling across institutions suggests a potential gap in knowledge exchange and collaboration, hindering the development of a unified research agenda. Strengthening connections between influential works could foster a more cohesive and integrated understanding of stakeholder opinion mining and sentiment analysis. By addressing these gaps and building on the contributions of existing research, scholars and practitioners can better navigate the complexities of crisis communication and bolster organizational resilience in an increasingly dynamic and uncertain environment.

References

Alpaslan, C. M., Green, S. E., & Mitroff, I. I. (2009). Corporate governance in the context of crises: Towards a stakeholder theory of crisis management. *Journal of Contingencies and Crisis Management*, *17*(1), 38-49.

Bosse, D. A., Phillips, R. A., & Harrison, J. S. (2009). Stakeholders, reciprocity, and firm performance. *Strategic management journal*, *30*(4), 447-456.

Degtjarjova, I., Lapina, I., & Freidenfelds, D. (2018). Student as stakeholder: Voice of customer in higher education quality development. *Маркетинг і менеджмент інновацій*(2), 388-398.

Freeman, R. E., Phillips, R., & Sisodia, R. (2020). Tensions in stakeholder theory. *Business & Society*, *59*(2), 213-231.

Larson, A. (1992). Network dyads in entrepreneurial settings: A study of the governance of exchange relationships. *Administrative science quarterly*, 76-104.

Mitroff, I. I., & Kilmann, R. H. (1984). *Corporate tragedies: Product tampering, sabotage, and other catastrophes*. Greenwood.

Molavi, H., & Zhang, L. (2024). Predicting the Next Frontier on Stakeholder Perception and Behaviour: Utilizing Bibliometrics to Identify Emerging External Influences on Management Studies. *Journal of Management Studies Annual Conference*,

Mwesigwa, R., Ntayi, J., Bagire, V., & Munene, J. C. (2018). Stakeholder behavior, relationship building practices and stakeholder management in Public Private Partnership Projects in Uganda.

Nathan, M. L., & Mitroff, I. I. (1991). The use of negotiated order theory as a tool for the analysis and development of an interorganisational field. *The Journal of applied behavioral science*, *27*(2), 163-180.

Nguyen, T. V., & Rose, J. (2009). Building trust—Evidence from Vietnamese entrepreneurs. *Journal of Business Venturing*, *24*(2), 165-182.

Pearson, C. M., & Clair, J. A. (1998). Reframing crisis management. *Academy of management review*, *23*(1), 59-76.

Perrow, C. (1999). *Normal accidents: Living with high risk technologies*. Princeton university press.

Rahimnia, F., & Molavi, H. (2021). A model for examining the effects of communication on innovation performance: Emphasis on the intermediary role of strategic decision-making speed. *European Journal of Innovation Management*, *24*(3), 1035-1056.

Savage, G. T., Nix, T. W., Whitehead, C. J., & Blair, J. D. (1991). Strategies for assessing and managing organisational stakeholders. *Academy of Management Perspectives*, *5*(2), 61-75.

Ulmer, R. R. (2001). Effective crisis management through established stakeholder relationships: Malden Mills as a case study. *Management communication quarterly*, *14*(4), 590-615.

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17803

Digital Transformation in Supply Chain Management: A Bibliometric Analysis

Lihong Zhang[1], Saeed Banihashemi[2], Aiwen Rui[1], Song Chen[3]

[1]Department of Civil Engineering and Management, University of Manchester, United Kingdom, [2]Faculty of Arts & Design, University of Canberra, Australia, [3]School of Economics and Management, Tongji University, PR China.

How to cite: Zhang, L.; Banihashemi Saeed.; Rui Aiwen.; Chen Song. 2024. Digital Transformation in Supply Chain Management: A Bibliometric Analysis. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17803

Abstract

In an era dominated by digital advancements, Supply Chain Management (SCM) is undergoing significant transformations. This study aims to outline a digitally-enabled SCM framework by examining prevalent research themes, methodologies, collaboration models, and groundbreaking contributions. Leveraging bibliometric analysis techniques, 600 articles from 2002 to 2023, sourced from 3 top databases, was conducted using CiteSpace for keyword analysis, co-citation analysis, and emerging term evaluations. The research identifies three distinct phases in this field: initial, incremental, and accelerated, with a notable surge in research activity post-2017, particularly after 2019. China, the US, and the UK are major contributors. Dominant topics include technology integration, efficiency, globalization challenges, strategic management, and sustainability. The study reveals limited collaboration among authors but highlights influential scholars. Given the centrality of DT, it emphasizes the need for interdisciplinary exploration and consideration of rapidly evolving digital paradigms in future research endeavors.

Keywords: *Digital Transformation; Supply Chain Management; Scientometric and Bibliometric Analysis; Digital Production; Industry 4.0.*

1. Introduction

Traditionally, SCM has been a specialized field focused on coordinating logistics, procurement, and operations (Preindl et al., 2020). However, the advent of the digital age has acted as a catalyst, fundamentally changing these traditional practices and generating an emerging paradigm of SCM 4.0. Importantly, this evolution goes beyond mere integration of digital tools; it marks a fundamental reimagining of the core nature of SCM. As disruptive technologies such

as blockchain, IoT, and machine learning become an integral part of modern SCM, it becomes imperative to critically assess their impact, challenges, and the future trajectory they chart (Zekhnini et al., 2022).

The advent of DT is instigating significant shifts across various facets of SCM, ranging from the intricacies of information acquisition and data analytics to the complexities of operational streamlining and sustainability imperatives (Dolgui et al., 2020). Hence, it is imperative to offer a holistic scrutiny of the manifold ramifications of DT within the domain of SCM, structured in alignment with the sequential phases of digitalization.

The integration of advanced technologies including AI, IoT, blockchain, and big data analytics in SCM has been a focus of recent academic research (Dolgui et al., 2020; Gomber et al., 2018). These technologies have enabled organizations to transition from traditional linear supply chains to more agile, transparent, connected, and responsive digital supply chain networks.

Using web text analysis and bibliometric techniques, the study aims to provide insights into research trends and guide future scholarly inquiry. By bridging the gap between academia and industry, the research offers practical guidance for organizations seeking to digitally transform their supply chain operations. Additionally, the study sheds light on the potential benefits and risks of digital supply chains, providing a framework for strategic planning and risk mitigation.

Employing CiteSpace software, the analysis aims to map the trajectory and hotspots in digital SCM research, identifying key research gaps and future areas of inquiry. Overall, the study contributes to a nuanced understanding of how digital technologies are shaping the landscape of SCM and provides valuable insights for both practitioners and academics in this rapidly evolving field.

2. Methodology Approach

This study employs bibliometric analysis as an interdisciplinary method to scrutinize and quantify scholarly literature, integrating text mining and network analysis techniques (Ball, 2017). The chosen tool, CiteSpace, is selected for its ability to simultaneously analyze data from multiple databases, including Web of Science (WoS), Scopus, and PubMed, providing a comprehensive panorama of scholarly literature.

The formulation of search criteria involved identifying key phrases and terms, incorporating advanced search functionalities across databases to accommodate keyword variability. The search query yielded a dataset of 600 articles from 2002 to July 31, 2023 [Figure1].

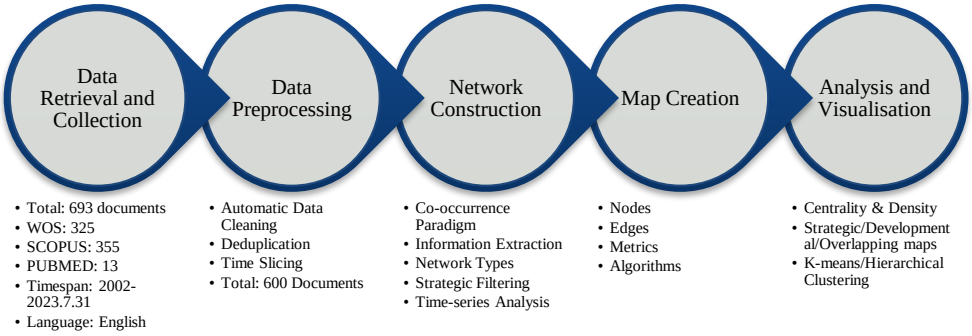


Figure 1. Workflow of the adopted bibliometric method

The specific search query employed is detailed in Table 1. Data preprocessing is conducted to address inconsistencies across different databases, including automated data cleaning, deduplication, and time-slicing options. CiteSpace is utilized for this purpose, providing advanced customization options for standardization. The final dataset of 600 articles serves as the foundation for subsequent bibliometric analysis in the study.

Table 1. Applied Search Query.

Database	Search Query	Result
WoS	((TS=("digital* transform*")) OR TS=(digitalization)) AND TS=("supply chain management") and English (Languages) and Article or Review Article (Document Types)	**325**
Scopus	(TITLE-ABS-KEY ("digital* transform*") OR TITLE-ABS-KEY-AUTH (digitalization)) AND TITLE-ABS-KEY ("supply chain management") AND (LIMIT-TO (DOCTYPE, "ar")) AND (LIMIT-TO (LANGUAGE, "English"))	**355**
Pubmed	(("digital transformation"[Title/Abstract]) OR (digitalization [Title/Abstract])) AND ("supply chain management"[Title/Abstract])	**13**

3. Results

3.1. Development stages of DT in SCM

The realm of DT research in SCM has undergone extraordinary expansion over the past two decades. Figure 2 features a red dashed line that best fits the growth trend of cumulative publications. An exponential fit model was found to provide the most accurate representation of this growth, with a Mean Square Error (MSE) of approximately 841.47. According to the exponential best-fit model, the cumulative number of articles is projected to reach approximately 651 by the end of 2023.

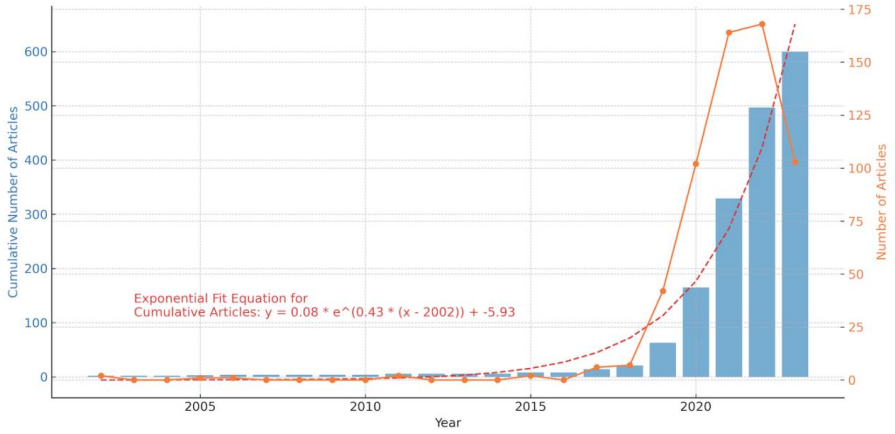


Figure 2. Literature trends of DT-SCM from 2002-2023

The holistic analysis indicates a marked shift occurs from 2017 onwards, with a significant uptick in annual publication volume. Rapid advancements in technologies pertinent to DT—such as the IoT and information technology—have likely been instrumental in fueling this increase in scholarly output. The year 2019 marks a pivotal moment, witnessing an exponential growth in publications. Exponential growth in a field typically signifies the influence of some underlying catalyst. In this context, the COVID-19 pandemic appears to be a plausible driver, as it likely accelerated DT initiatives. The pandemic necessitated shifts to remote work and amplified the importance of online business operations. Consequently, it exerted considerable impact on global supply chains, spurring increased research and practical applications in DT (Khan et al., 2023). Furthermore, since 2019, the rapid popularization of AI and big data technologies has deepened both corporate and academic understanding of DT's significance. This has caused digitalization to evolve from a mere technological trend to a market imperative, giving rise to concepts like agile SCM and digital SCM (Centobelli et al., 2020). In summary, the field of SCM has increasingly focused on leveraging cutting-edge technologies, elevating DT from a peripheral subject to a central area of scholarly and practical focus.

3.2. Keywords mapping

Keywords serve as a succinct summary of a paper's primary focus, making keyword analysis a valuable tool for identifying trending topics within a specific field. Figure 3 displays a keyword co-occurrence map generated by CiteSpace, focusing on DT research in SCM. In this figure, each circular node represents a keyword, with larger nodes denoting higher keyword frequencies. The links between nodes indicate the degree of relational closeness between respective keywords. The network comprises 403 nodes and 1903 edges, representing 403 individual keywords and their co-occurrence relationships. With a network density of 0.0235, the map suggests a relatively close degree of co-occurrence among the keywords. The largest

connected component encompasses 367 nodes, accounting for 91% of the total, indicating a strong core group of interrelated keywords. Flagged nodes, which make up 1.0% of the total, likely represent core concepts, technologies, or methodologies in the field, such as "big data analytics", "sustainable SCM", "performance" and "information technology". "Performance", "impact", and "challenge" have emerged as prominent themes in recent research, while "model" and "logistics" are underrepresented despite their central roles. Keywords like "internet" and "innovation" may represent nascent areas for exploration.

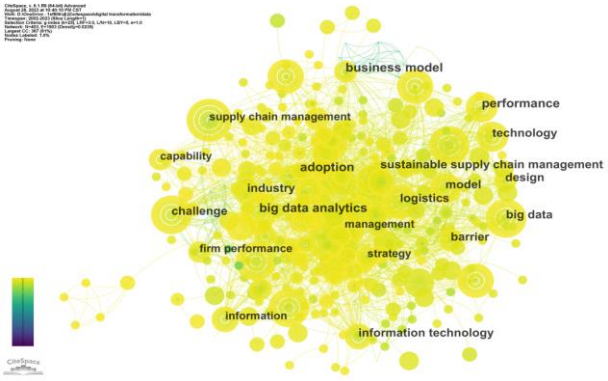


Figure 3. Keywords co-occurrence network of DT-SCM

3.2.1. Keywords Burst

The analysis of keyword bursts highlights evolving foci and thematic trajectories in the realm of DT in SCM. Keywords such as "research agenda" from 2017 to 2020 signify the strategic importance of DT in SCM, reflecting a shift towards applied research aimed at optimizing supply chains. The term "Internet of Things" saw exceptional intensity in 2020, emphasizing its crucial role in shaping contemporary supply chain systems. Emergent keywords like "green" and "sustainability" indicate a growing awareness of environmental stewardship in SCM, suggesting a potential increase in research on the environmental efficacy of DT strategies. Keywords like "model" and "logistics" exhibit high centrality metrics, underscoring their pivotal yet underexplored roles in the academic landscape of the field and warranting further scholarly attention.

3.2.2. Keywords cluster

Figure 4 highlights strong breakout words that have shown marked prominence across the entire dataset, independent of temporal considerations. Due to the abundance of keywords, the Log-Likelihood Ratio (LLR) algorithm was employed to cluster keywords that either have similar meanings or are used to describe comparable concepts. In the domain of DT-SCM, keywords are organized into 10 distinct clusters, as depicted in Figure 5. In this representation, cluster #0 is the largest, followed by cluster #1, and so on. The map reveals that all clusters exhibit a degree

of overlap, suggesting a close interrelationship among them. This is particularly true for the first seven clusters, which are almost entirely overlapping.

Cluster #0 centers on "management efficiency" and the "digital economy"; Cluster #1 deals with "technology adoption and decision"; Cluster #2 focuses on "sustainable supply chain management"; Cluster #3 primarily examines the impact of the "COVID-19 pandemic" on SCM; Cluster #4 emphasizes "bibliometric analysis" in the supply chain; and Cluster #5 pertains to "industry 4.0" and the "global value chains". Among these, Cluster #0 and Cluster #1 are the largest and are primarily concerned with technology adoption and SCM efficiency, marking them as current areas of intense research focus. The highest-quality cluster, Cluster #10, with a Silhouette value of 0.935, underscores that the topic of the "development model" is both highly specific and well-defined. For instance, Cluster #0 clearly underscores the impact of DT on enhancing management efficiency within SCM. Cluster #2 reveals an integration of emerging technologies into SCM practices, starting from around 2018. The year 2020 marked a pivotal moment, as the advent of the COVID-19 pandemic prompted significant shifts in business models and accelerated the reliance on digital means, thereby propelling the DT of supply chains. Within Cluster #5, the entry into the Industry 4.0 era in recent years has had a notable impact on supply chain models.

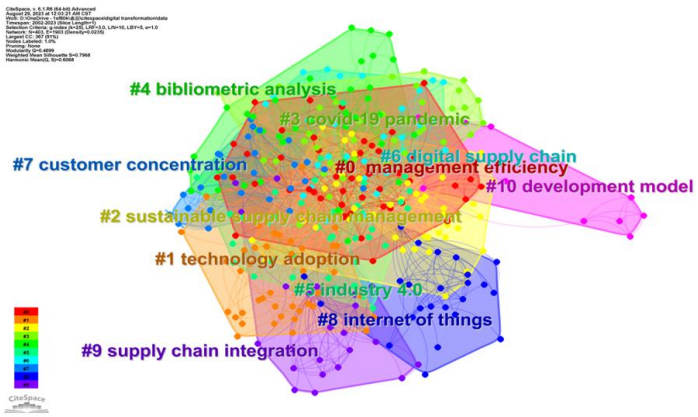


Figure 4. Keywords cluster of DT-SCM

3.3. Keywords evolution analysis

Keyword timezone analysis provides a temporal overview of research evolution in Digital Transformation within Supply Chain Management (DT-SCM). In Figure 5, research on SCM traces back to 2005, with early emphasis on business models. A hiatus in significant activity is observed until 2017 when applied research merges science and technology with SCM, particularly focusing on information technology. By 2019, with the advent of big data and

Industry 4.0, research in DT-SCM experiences an explosive surge, addressing challenges and barriers associated with DT.

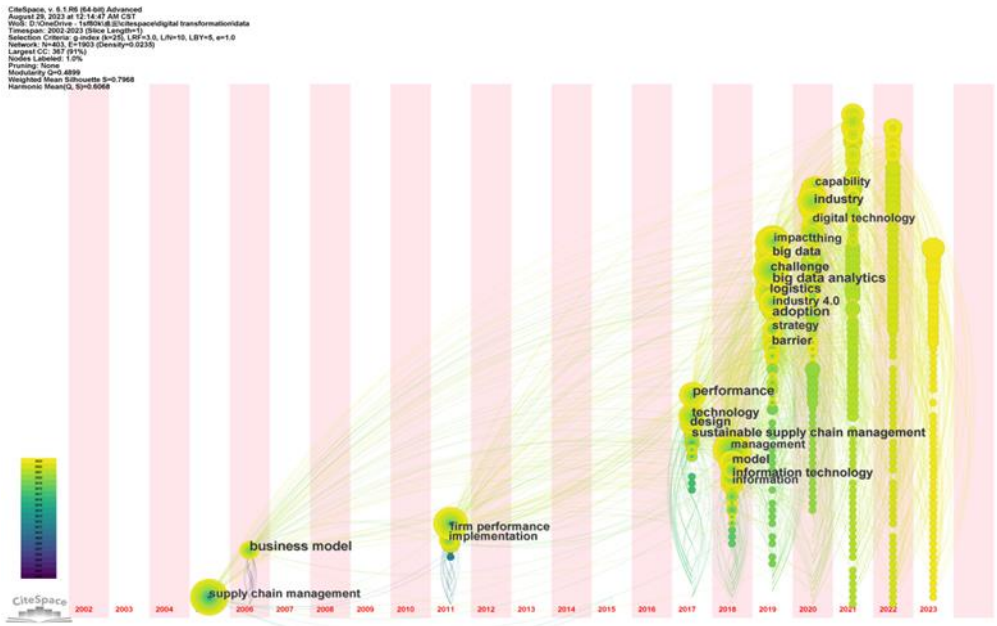


Figure 5. Keyword timezone of DT-SCM

Further examination in Figure 6 reveals intensified cross-linkages among research foci post-2017, emphasizing the need for a holistic evaluation of interplay between thematic clusters. For instance, the relationship between management efficiency and technological adoption signifies the role of innovations in supply chain efficiencies. Similarly, connections between Industry 4.0 and the COVID-19 pandemic highlight the potential applications of technology in emergency response frameworks, emphasizing the need for multifaceted solutions to supply chain challenges.

4. Discussion

This section synthesizes the findings of the scientometric and bibliometric analyses with the literature review, highlighting the rapid ascent of Digital Transformation in Supply Chain Management (DT-SCM) and outlining future research directions.

Determinants of DT in Traditional Supply Chain Structure: Highly cited literature and authors offer insights into influential perspectives, emphasizing that DT extends beyond technology to address global SCM challenges. The surge in publications on DT research within SCM, particularly post-2017, aligns with the advent of Industry 4.0 and the exponential growth of technologies integral to DT.

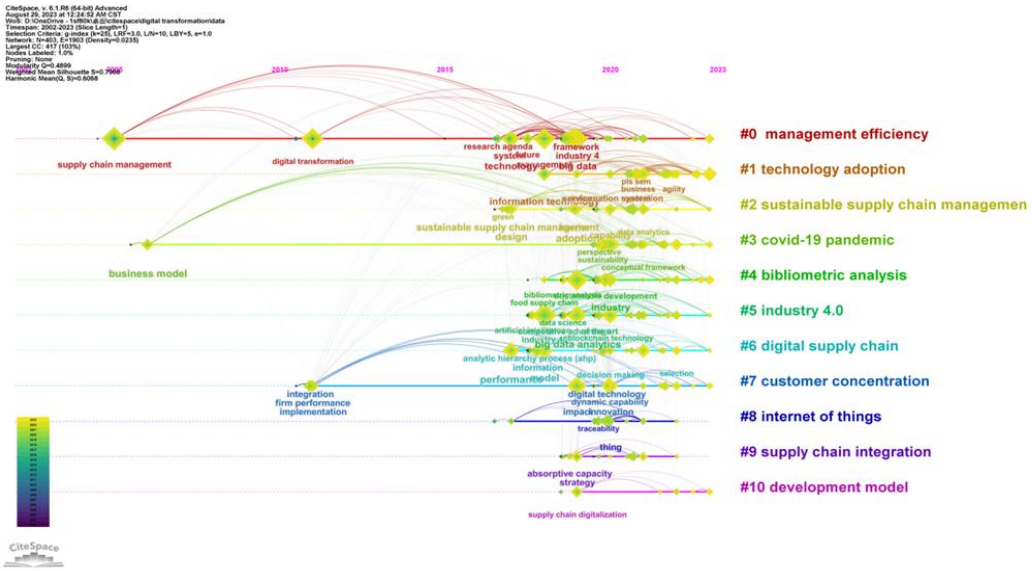


Figure 6. Keywords cluster timeline on DT-SCM

Current Status and Research Focus of DT-SCM: Keyword analysis reveals hotspots and core concepts, with emerging themes like big data analytics and information technology gaining prominence. Sustainable SCM emerges as a critical area of investigation, reflecting a paradigmatic shift towards environmental stewardship in digital SCM. While management efficiency remains a dominant cluster, research increasingly integrates avant-garde technologies like AI and big data.

Future Research Trends and Directions of DT-SCM: The evolution of keywords over time suggests a trajectory towards sustainable SCM and IoT-driven applications. Multidisciplinary research will likely tackle challenges and nuances associated with digital SCM, emphasizing the integration of digital tools and sustainability initiatives. The discussion also underscores the necessity for interdisciplinary collaboration and the exploration of emergent themes.

DT-SCM Evolution Model: The developed framework provides a strategic tool for navigating the complex landscape of DT in SCM. It encompasses the current state of research, research challenges and limitations, future trajectories, determinants of DT-SCM, influential perspectives, and emerging themes. The model underscores the importance of addressing research gaps, leveraging influential perspectives, and considering the broader implications of technological advancements in SCM [Figure 7].

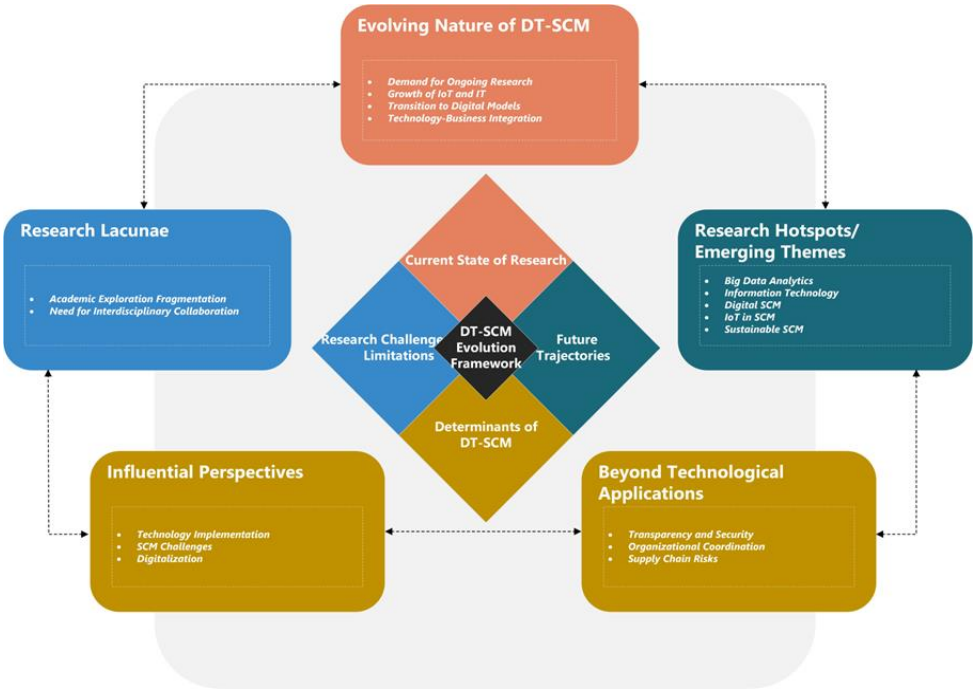


Figure 7. DT-SCM evolution framework

References

Ball, R. (2017). An introduction to bibliometrics: New development and trends. Chandos Publishing.

Centobelli, P., Cerchione, R., & Ertz, M. (2020). Agile supply chain management: where did it come from and where will it go in the era of digital transformation? *Industrial Marketing Management*, 90, 324-345.

Dolgui, A., Ivanov, D., & Sokolov, B. (2020). Reconfigurable supply chain: The X-network. International Journal of Production Research, 58(13), 4138-4163.

Gomber, P., Kauffman, R. J., Parker, C., & Weber, B. W. (2018). On the fintech revolution: Interpreting the forces of innovation, disruption, and transformation in financial services. *Journal of management information systems*, 35(1), 220-265.

Khan, O., Huth, M., Zsidisin, G. A., & Henke, M. (2023). Supply Chain Resilience: Reconceptualizing Risk Management in a Post-Pandemic World. Springer.

Preindl, R., Nikolopoulos, K., & Litsiou, K. (2020). Transformation strategies for the supply chain: The impact of industry 4.0 and digital transformation. *Supply Chain Forum: An International Journal*.

Zekhnini, K., Cherrafi, A., Bouhaddou, I., Chaouni Benabdellah, A., & Bag, S. (2022). A model integrating lean and green practices for viable, sustainable, and digital supply chain performance. *International Journal of Production Research*, 60(21), 6529-6555.

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17833

A scientometric review on green manufacturing systems for small and medium sized enterprises (SMEs)

Jorge Naranjo Perez[1], Lihong Zhang[2], Xirong Li[3]

[1]Department of Mechanical, Aerospace and Civil Engineering, University of Manchester, United Kingdom, [2]Department of Project Management, University of Manchester, United Kingdom, [3]Department of Mechanical, Aerospace and Civil Engineering, University of Manchester, United Kingdom.

How to cite: Naranjo Perez, J.; Zhang, L.; Li, X. 2024. Paper A scientometric review on green manufacturing systems for small and medium sized enterprises (SMEs). In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17833

Abstract

This paper will discuss the current state, barriers, and potential opportunities that small and medium-sized enterprises face in an ever-changing need for sustainable innovation in manufacturing industrial processes. A scientometric literature review has been conducted to provide a solid understanding of green manufacturing systems, specifically on developing manufacturing businesses which may lack intellectual capital, resources and technology as opposed to multinational settled corporations. This literature review also examines potential advantages, drawbacks, and solutions for those SMEs who adopt these sustainable and green manufacturing practices. The employment of the software CiteSpace has aided to deliver this methodology and demonstrate a bigger picture of the prevailing scientific literature that is currently available. The key findings reveal a range of challenges faced by SMEs depending on their geographical positioning. A step-by-step approach beginning with the implementation of cost-efficient management techniques and advancing towards investments in green technologies can facilitate this transition.

Keywords: *Scientometrics; Green manufacturing; Small and medium sized enterprises (SMEs).*

1. Introduction

The importance of manufacturing systems is to produce goods with a given input. This input includes raw materials, machinery, and workforce. They all work together to produce finished goods which can satisfy customer demand (Nassimbeni, 2003). The manufacturing industry has a very strong correlation with the economic growth of a country, and it is the driver of productivity development, according to Speering (2018). The need for sustainability

implementation in manufacturing systems has been essential due to multiple factors such as its implications on the environment; tougher government regulations; soaring energy prices and minimization of waste to reduce overhead costs (Giret et al., 2015).

The current research gap to be revised is that SMEs together account for almost 70% of the global pollution (Gurria, 2018), hence thorough scrutiny must be taken into this topic to contribute towards a more environmentally friendly planet. The reason SMEs have such a strong impact on this given data is that they account for 90% of businesses worldwide; influence a quarter of the world's population and supply larger companies with the relevant resources (Shrimali, 2022).

Due to the ever-increasing customer demand towards eco-friendly manufacturing business strategies, it is essential for SMEs to tackle these by exploring ways of waste minimization, reducing energy consumption, and adopting circular economy principles. These include cost-effectively adopting renewable energy sources, recycling, and reusing materials by implementing environmentally safe production processes to reduce energy consumption and toxic waste footprints. The following graphs show a clear illustration of the density number of micro, small and medium sized enterprises (MSMEs) in the world, split into different colour codes (see Figure 1).

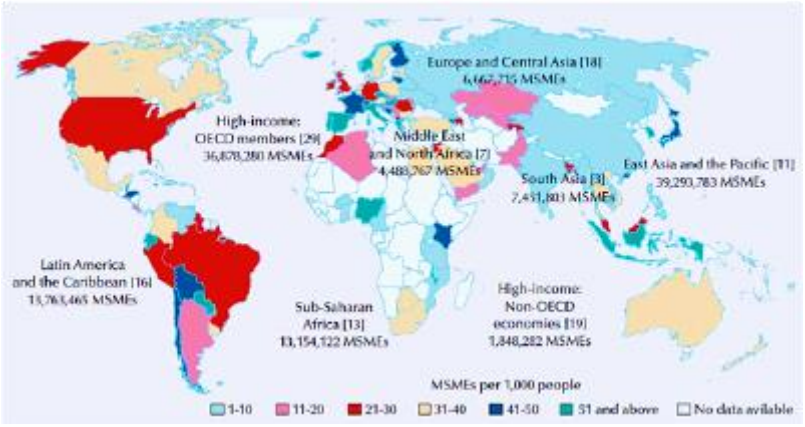


Figure 1. MSMEs density around the world, (Robu, 2013).

Countries with a green colour code, have the greatest number of MSMEs per 1,000 habitants. Based on Giret et al. (2015) previous statement, we can deduce that countries such as Spain or Indonesia, have highly dependent economies on MSMEs. In terms of actual numbers of MSMEs, East Asia and the Pacific have the highest number of MSMEs, with 39,293,783 of them. This serves as a demonstration of the imperative need to meticulously address the environmental footprint of SMEs, given their significant influence on a worldwide scale.

Additionally, these emission practices are categorized in the Scope 3 emissions (Shrimali, 2022). Scope 3 emissions refer to pollutants that occur as a result of a company's activities but are not directly owned by the company. This includes emissions from sources such as waste disposal of purchased goods and services from suppliers or employee commuting via public transportation (Li et al., 2019). This particular type is 11.4 times greater than normal emissions and is frequently the most difficult to eliminate in comparison with scopes 1 and 2 (Shrimali, 2022).

Study shows that the main challenges faced by SMEs in the adoption of sustainable practices are access to funds, advanced technological assistance, and a motivated workforce. It is clear to say that SMEs in economically strong countries have greater ease in the application of these compared to LEDCs such as India, Thailand, Indonesia, etc (Abdullah et al., 2023). This statement can be strongly supported by the fact that developed countries have a carbon dioxide footprint four times above, compared to developing nations and threefold energy intensity (World Bank, 2013), which means that there is a clear discrepancy depending on how well economically developed the specific country is, where the SME is settled.

The aim of this paper is to make a thorough examination of the existing literature concerning green manufacturing systems for SMEs, while also explore potential research topics for future investigation which will guide practical research efforts and minimize the environmental footprint caused by manufacturing systems.

The primary objective is to examine a range of strategies, technology implementations, and practices that enable SMEs to reduce their environmental impact, and simultaneously, enhance their operational efficiency. Nevertheless, a robust awareness has to be made exposing the various barriers and opportunities SMEs face in adopting these methods to influence policy makers and encourage these implementations through appropriate recommendations. Despite these challenges, green manufacturing offers countless benefits which contribute towards the common goal of promoting a greener and more sustainable future. A scientometric approach on the current literature depth in this field will be exposed with the use of the software CiteSpace. This will help visualize trends and citation impacts between numerous scientific publications.

2. Methodology

Scientometrics is the mensurable study of scientific research. It involves the evaluation of scientific publications, citations, collaborations, and impact factors to provide an understanding into trends of scientific research on a particular field (Masic, 2022). The software program that will be used is CiteSpace due to its advanced features, that allow a high volume of data processing, tutorial systems, and user forums to provide necessary help, as well as a range of interactive tools including size and colour combinations (Chen, 2006). A software comparison to obtain the most optimal platform was done against VOSviewer, however, this software lacks

advanced features for new research fields, has limited functionality with respect to CiteSpace, and is slower in processing large datasets (Ding & Yang, 2020). In the following, a 4-step process will be presented that will show how the input data will be collected, introduced in the program, and analysed.

2.1. Data collection

A dataset of articles was obtained from the Web of Science by selecting the following search strings: [‘green manufacturing systems’ or ‘sustainable manufacturing systems’ and ‘SME’ or ‘small enterprises’]. The searches were limited to topic category only. This means that articles were only selected if the given strings were located either in the title, abstract or author keywords. This created a more accurate approximation for the extraction of admissible papers. After obtaining the dataset, a screening approach was implemented to enhance the refinement of papers for the literature review. The subsequent systematic process of selection and screening is depicted in Figure 2.

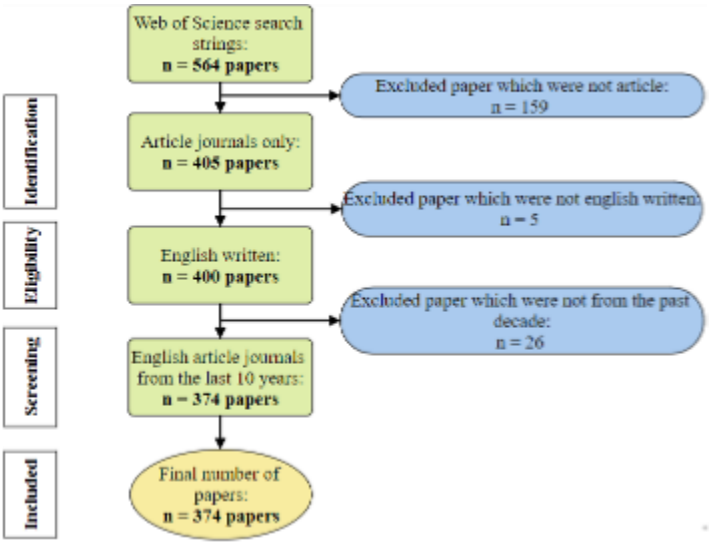


Figure 2. Self-adapted systematic review flow chart (Nandi and Nedumaran, 2021)

The search engine identified a total of 564 papers containing the relevant key terms shown in Figure 3. However, to narrow down the search, the selection was refined to include only article types and those that were English written. This led to a significant downsize of 416 papers. The number of papers published per year using this refinement is shown in the following bar chart and adapted from the Web of Science results (*see Figure 3*).

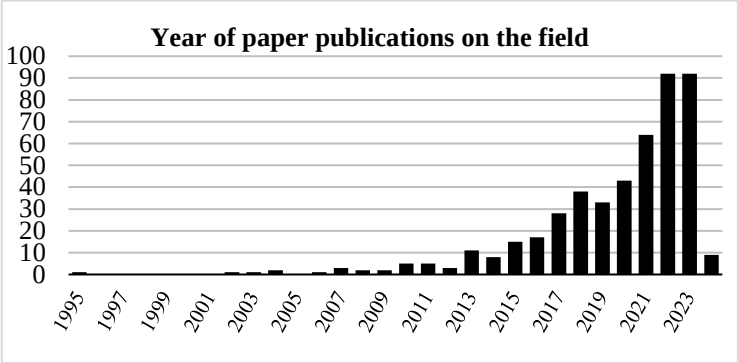


Figure 3. Bar chart showing number of publications from 1995 to 2024.

An exponential growth of article papers relating green manufacturing systems for SMEs can be seen in the past decade. A consideration of data spanning in the past decade was selected, as it comprised a 92.38% of all results, thereby enabling an accurate sample of papers to be reviewed. This resulted in 374 article papers that will be utilized for the science mappings.

3. Results and Findings

3.1 Networks

A mesh-like structure was generated giving valuable information about co-citation data, citation bursts and result visualizations. This specific type of network has been separated into clusters, shown in Figure 4.

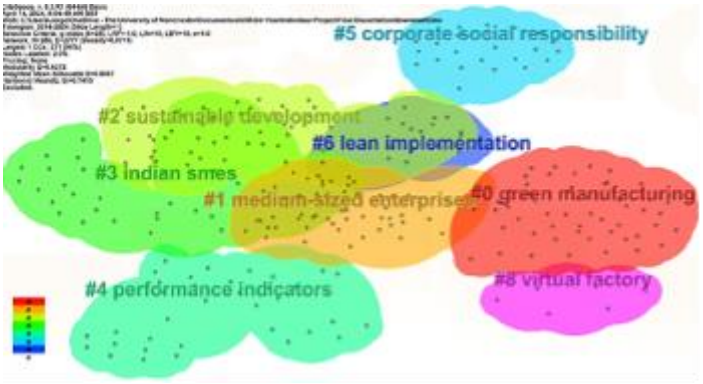


Figure 4. Cluster Network generated by CiteSpace.

A cluster refers to a group of related scientific publications that are closely linked based on citation patterns. These cluster visually represents the relationship between different papers (Chen, 2006). The tiny circle connections are nodes, which represent individual articles. Node

spacing is influenced by factors number of citations and overall network density (Liu et al., 2022). The next meshwork generated shows a co-citation mapping, showing highly cited papers.

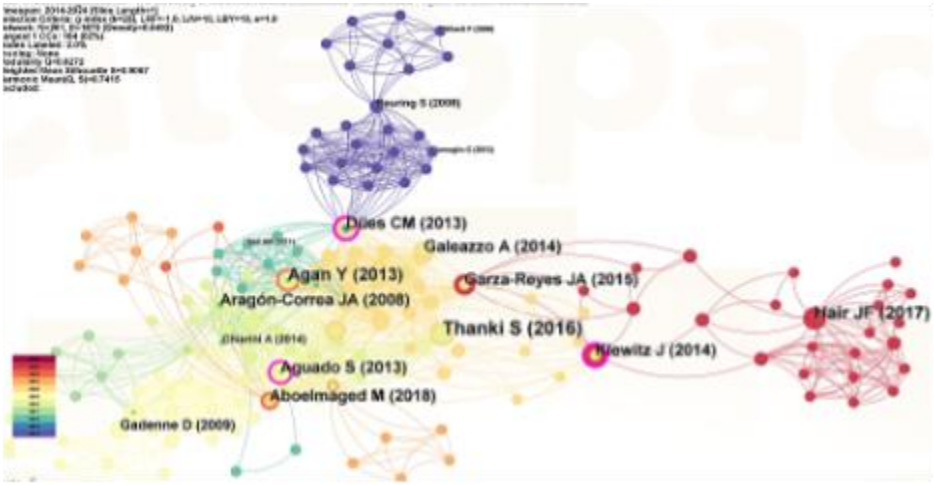


Figure 5. Co-citation network generated by CiteSpace.

Additionally, the institution and country network map highlight the global distribution collaboration based on co-authorship association. It helps identify connections between different organizations and regions and classify potential research patterns, emerging collaborative opportunities, or world scientific sequences (Zheng and Wang, 2019).

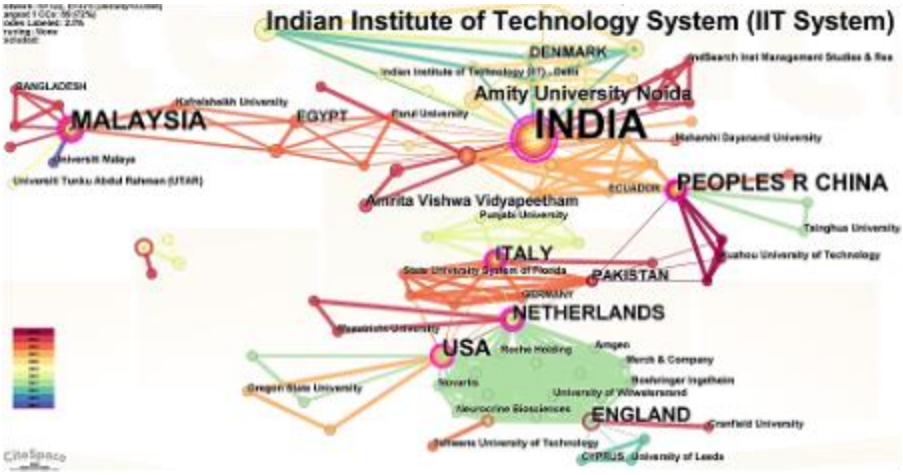


Figure 6. Institution and country network.

3.2. Analysis of the output network

In the cluster network diagram, 8 observable clusters can be identified. There are 4 clusters which are highly concentrated in the middle of the diagram '*sustainable development*', '*medium sized enterprises*', '*lean implementation*', and '*Indian SMEs*'. This tells that a great effort and scientific research has been done in those fields.

Through the co-citation network from Figure 5, it can be seen that Hair F. has been emerging actively during recent years; however it has been highly diverged from the middle of the map, which may suggest that low collaboration has been done with previous work from other authors. This can be interpreted as a beneficial development, as he is departing from the established scientific research and venturing into unexplored realms of study.

Those countries that have made a significant advancement in this field are India, followed by Malaysia, USA, and China respectively. This is supported by their node size and high centrality degree based on their ring size. A higher degree of centrality is observable in developing countries as opposed to economically developed countries which may signify a greater need of scientific research in those areas. The Indian Institute of Technology System (IIT System) has been the biggest research institution on this matter and substantial collaborative work can be seen based on their high linkage degree. Nodes extending from China by the Xuzhou University of Technology is indicative of their active publication of papers and ongoing advancements in research. This is further supported by their delineation and colour coding on the map, pointing towards developments made in the year 2024.

4. Conclusion

Green is an integral component of the foundational principles of sustainability, in conjunction with the economic and social pillars. To ensure a sustainable machining process, green manufacturing systems must be fulfilled by effectively reducing waste and energy consumption through a combination of management and technological strategies. SMEs with limited resources can begin implementing cost-effective management approaches by leveraging circular economy strategies. Once these approaches have been successfully integrated, SMEs may consider investing in technological solutions including efficient production machinery, renewable energy utilization, and environmental management systems. A notable focus of research centrality has been discovered in India, which highlights the necessity to increase scientific endeavours to offer valuable insights and methodologies for SMEs to effectively transition towards sustainable green manufacturing practices.

In conclusion, it is hoped that this study has provided valuable information for authors seeking to embark into this topic and understand the direction and process involved. The integration of green manufacturing systems in SMEs is imperative due to their worldwide influence in

providing a prosperous future for not only the economy but preserving the existing natural ecosystems for future generations.

References

Abdullah, A., Saraswat, S., & Talib, F. (2023). Barriers and strategies for sustainable manufacturing implementation in SMEs: A hybrid fuzzy AHP-TOPSIS framework. *Sustainable Manufacturing and Service Economics*, *2*, 100012. https://doi.org/10.1016/j.smse.2023.100012

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, *57*(3), 359–377. https://doi.org/10.1002/asi.20317

Ding, X., & Yang, Z. (2020). Knowledge mapping of platform research: a visual analysis using VOSviewer and CiteSpace. *Electronic Commerce Research*, *22*. https://doi.org/10.1007/s10660-020-09410-7

Giret, A., Trentesaux, D., & Prabhu, V. (2015). Sustainability in Manufacturing Operations scheduling: a State of the Art Review. *Journal of Manufacturing Systems*, *37*, 126 140. https://doi.org/10.1016/j.jmsy.2015.08.002

Gurría, A. (2018). SMEs are key for more inclusive growth. *OECD Observer*. https://doi.org/10.1787/25aa3d56-en

Li, M., Wiedmann, T., & Hadjikakou, M. (2019). Enabling Full Supply Chain Corporate Responsibility: Scope 3 Emissions Targets for Ambitious Climate Change Mitigation. *Environmental Science & Technology*. https://doi.org/10.1021/acs.est.9b05245

Liu, D., Che, S., & Zhu, W. (2022). Visualizing the Knowledge Domain of Academic Mobility Research from 2010 to 2020: A Bibliometric Analysis Using CiteSpace. *SAGE Open*, *12*(1), 215824402110685. https://doi.org/10.1177/21582440211068510

Masic, I. (2022). Scientometrics: The imperative for scientific validity of the scientific publications content. *Science, Art and Religion*, *1*(1), 56–80. https://doi.org/10.5005/jp-journals-11005-0017

Nandi, R., & Nedumaran, S. (2021). Understanding the Aspirations of Farming Communities in Developing Countries: a Systematic Review of the Literature. *The European Journal of Development Research*. https://doi.org/10.1057/s41287-021-00413-0

Nassimbeni, G. (2003). Local manufacturing systems and global economy: Are they compatible?: The case of the italian eyewear district. *Journal of Operations Management*, *21*(2), 151–171. https://doi.org/10.1016/S0272-6963(02)00090-6

Robu, M. (2013). The dynamic and importance of smes in economy. *The USV Annals of Economics and Public Administration*, *13*, 84–89. https://ideas.repec.org/a/scm/usvaep/v13y2013i1(17)p84-89.html

Shrimali, G. (2022). Scope 3 emissions: Measurement and management. *The Journal of Impact and ESG Investing*. https://doi.org/10.3905/jesg.2022.1.051

Speering, J. (2018, July 29). *Why manufacturing is essential for economic growth?* Aspioneer. https://aspioneer.com/why-manufacturing-is-essential-for-economic-growth/

World Bank. (2013). *World Development Indicators 2013.* World Bank.

Zheng, K., & Wang, X. (2019). Publications on the Association between Cognitive Function and Pain from 2000 to 2018: a Bibliometric Analysis Using CiteSpace. *Medical Science Monitor*, *25*, 8940–8951. https://doi.org/10.12659/msm.917742

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17841

The confluence of project and innovation management: Scientometric mapping

Lihong Zhang[1], Saeed Banihashemi[2], Yujue Zhang[1], Song Chen[3]

[1]Department of Civil Engineering and Management, University of Manchester, United Kingdom, [2] Faculty of Arts & Design, University of Canberra, Australia, [3]School of Economics and Management, Tongji University, PR China.

How to cite: Zhang, L.; Banihashemi, S.; Zhang, Y.; Chen, S. 2024. The confluence of project and innovation management: Scientometric mapping. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17841

Abstract

The convergence of research between innovation management (IM) and project management (PM) has been increasingly noted. To address and reflect this rapid intersection, this study conducts a visualised bibliometric review of 521 articles from 2003 to 2023, sourced from WOS, Scopus, and PubMed. Through publication metric analysis, disciplinary distribution, collaborative networks, and keyword mappings, research synergies and landmarks are identified. Academic advancements, dominant research themes, and frontier fields within the domain are recognised. This pioneering cross-disciplinary exploration offers insights for industry professionals and researchers. Key findings include predominant subjects (management, engineering, and business), significant research landmarks (Stage-Gate system, dynamic capabilities), dominant research themes (innovation initiatives, methodologies, practical applications), and emerging frontier fields (artificial intelligence, agile product management, new product development approaches). A three-stage evolution framework of PIM is proposed, aiding in understanding managerial and organisational adaptations amidst technological and societal changes.

Keywords: *Bibliometric Review; Innovation Management; Project Management; Scientometric Mapping*

1. Introduction

Innovation management (IM) and project management (PM) are intrinsically linked and intertwined, requiring a blend of market insights and technical expertise within structured frameworks(Silva and Gil, 2013). Projects now extend beyond operational enhancements to encompass new product development, entrepreneurship, and strategic initiatives (Davies, 2014).

The adoption of PM is driven by the imperative to identify success factors, especially amidst industry competition and shifting consumer preferences (Honorato and De Melo, 2023). To navigate these challenges, organisations increasingly integrate innovative PM strategies to maintain market competitiveness (Young et al., 2012). However, managing innovative projects entails complexity and risk, necessitating careful management to avoid adverse outcomes (Pinto et al., 2011).

2. Background

2.1. Project Management

The APMBOK Guide (APM, 2019) defines PM as a temporary endeavour centred on creating unique products or services and meeting stakeholder expectations, highlighting it as a performance-driven discipline, effectively organising and managing project activities.

PM's importance is growing in both academic and organisational contexts, particularly in today's challenging economic environment (Oliveira Lucena et al., 2019). It has evolved significantly since the mid-20th century, transitioning from case-specific methodologies to standardised approaches applicable across various complex sectors such as defence, construction, and IT (Davies, 2014). Despite comprehensive standards, studies indicate suboptimal PM practices, driving organisations to explore innovative strategies to enhance project success (Khalife et al., 2021). Traditional PM, often based on predictable models, may struggle to adapt to changing economic and business needs (Morris, 2013). Innovative projects require flexible strategies to adapt to unexpected challenges (Davies, 2014).

Given the uncertainties and complexities inherent in innovative environments, traditional approaches often fall short. This has led to the development of new theories and practices. The 'optimisation school' (Lenfle and Soderlund, 2019), design thinking (Ben Mahmoud-Jouini et al., 2016), and agile PM methods (hereafter agile unless otherwise stated) (Young et al., 2012) are prominent examples of these new approaches, aimed at enhancing adaptability and responsiveness in PM.

2.2. Innovation Management

The definition of innovation in the third edition of the Oslo Manual (Gault, 2013), highlights innovation extending beyond products to include various organisational processes. From a macro perspective, innovation is a transformative process where an advanced product or new process replaces its predecessor. Realising these innovations requires financial commitment and knowledge integration (Guerra Betancourt et al., 2013), making an innovation project both a transformative journey and an innovative venture, potentially leading to significant outcomes and pioneering solutions.

Innovation studies span various academic fields, focusing on uncertainties in developing and commercialising new products, processes, or services (Dodgson and Gann, 2011). It is crucial for businesses to thrive in ever-evolving technological and market environments (Goldhar, 1994). Research often involves contingency theory and organisational design, exploring how organisations adapt to uncertainty, complexity, and change. Projects or matrix structures are effective in overcoming these challenges (Mentzer, 1987). Moreover, organic organisational structures, known for their flexibility, are deemed conducive to innovation (Burns and Stalker, 1994).

In PM, innovation is often underrepresented in mainstream literature due to the differences between innovative and traditional projects (Tomala, 2004). Innovative approaches, dealing with uncertainties and complexities, contrast with traditional approaches focused on implementing existing decisions (Russo et al., 2017). Innovation in projects can be categorised as incremental, radical, or intermediate, correlating with derivative, breakthrough, and platform projects (Wheelwright and Clark, 1992). Various management strategies have been proposed to handle these types of projects. Ansoff, (2007) suggests managing proactive and reactive expectations in innovation projects, while Bibarsov et al., (2017) advocate combining long-term management tools with scientific principles such as selective management and goal orientation. Additionally, Shenhar and Dvir, (2007) proposed an adaptive PM model to enhance innovation and manage VUCA (volatility, uncertainty, complexity and ambiguity) challenges in a highly turbulent environment (Bennett, 2014).

2.3. Confluence of innovation and project management

Theoretical connections between Project and Innovation Management (PIM) have been explored, revealing a growing exchange of ideas in the twenty-first century. Scholars argue that innovation and contemporary PM are inherently linked, with projects often driving innovation in organisations (Silva and Gil, 2013; Davies, 2014). The literature on PM in innovation scenarios has evolved to include diverse theoretical bases, such as the PM paradigm, contingency theory, and organisational perspectives (Morris, 2013; Shenhar and Dvir, 2007).

Initially, the literature on PM and IM followed a separate and fairly self-contained trajectory of theoretical and professional growth (Davies, 2014), but recent trends indicate a convergence of ideas. Scholars are turning to interdisciplinary approaches that concentrate on how organisations deal with and manage innovation projects' uncertainty. Consequently, there is a clear research gap in project innovation, with a notable absence of comprehensive reviews consolidating and critically assessing existing studies in this intersecting domain.

Bibliometric analysis serves as a recognised method for surveying and summarising previous research, identifying academic trends, and predicting future research directions in PM (Silvius, 2017). It has been utilised to investigate various subfields of PM, including knowledge

management, project complexity, and project sustainability (De Rezende et al., 2018) , encompassing areas such as large-scale projects, construction initiatives, and software development (Lechler and Yang, 2017; Utama et al., 2020). However, in the era of big data, an econometric literature review offers a valuable approach to cross-integrate potentially connected disciplines such as PM and IM.

Existing research in the field of IM primarily focuses on models across various industries, including innovation projects in manufacturing, open innovation in pharmaceuticals, and IM models in aerospace (Honorato and De Melo, 2023). Although there are bibliometric reviews of IM studies covering evolution, models, techniques, and professionalisation (Robbins and O'Connor, 2023), there has yet to be a comprehensive literature review addressing the project-based context within IM, indicating a notable gap in current research.

2.4. Research Gaps and Objectives

Existing literature reviews often treat PM and IM as separate subjects, overlooking their potential intersections. While some reviews explore PM and IM individually, systematic examinations of their convergence, especially in the context of innovation and PM, are lacking. This gap persists despite technological shifts and societal changes spanning decades. To address the identified research gap, this study aims to elucidate the convergence within the PIM domain by examining publications from the past two decades. Specifically, the study sets out to accomplish the following objectives: (1) Provide an overview of the 20-year evolution of the PIM domain, emphasising publication statistics and disciplinary distribution. (2) Recognise research landmarks with highly-cited references and authors. (3) Discover the evolution of research advancements, the dominant research themes, and the frontier research fields by a series of keywords analysis.

3. Review Methodology

This study adopts bibliometric analysis, employed as a quantitative research method, assesses published literature within a specific knowledge domain (Abbasi et al., 2011) with scientometric analysis, complemented by visual mapping, offers a robust, replicable, and adaptable technique for tracing emerging trends and pinpointing pivotal contributions in a field (Chen et al., 2012). The data analysis software CiteSpace was selected for this review due to its robust mining and data compatibility processing capabilities (Zhang et al., 2023).

The bibliometric search held in three databases which are Web of Science (WOS), Scopus, and PubMed. To assure the accuracy of the literature scope, this study used a query-based search method (The search query for title, abstract and key words: ("project management" OR "project governance") AND ("innovation management" OR "innovation project*") OR ("innovative project" OR "project innovation")) to conduct a preliminary scoping search in the database

which held in August 31, 2023, as the time point and the accumulation of results yielded 1143 valid literature information sets. Then we applied our inclusion & exclusion criteria. (1) for the quality purposes, only journal articles were included and book reviews, editorials, and conference papers were excluded. (2) The time frame was limited to 2003 to 2023, as the search results indicated that most journal articles were published after 2003. (3) Only journal articles published in English were incorporated. This screening process end up with a total of 573 articles. Figure 1 illustrates the core methodologies employed in this research, detailing the data extraction process.

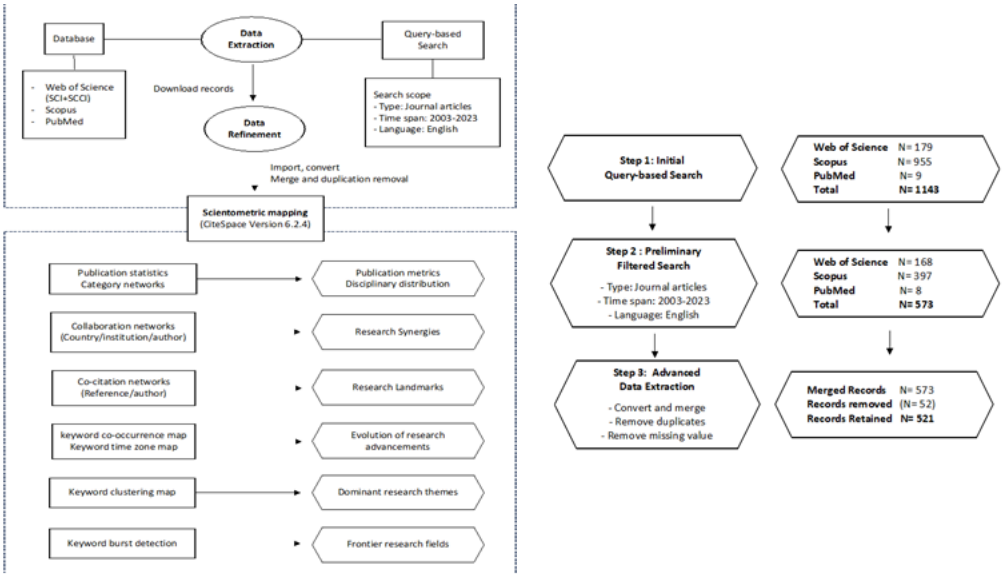


Figure 1. Review methodology (Left) and Process of data extraction (Right) (Authors' Own Source)

3.1. Procedure in CiteSpace

A total of 573 records were obtained and imported into CiteSpace for file format conversion, data merging, elimination of duplicates, and removal of records with missing values. After further data cleansing, 521 bibliographic records were retained for scientometric analysis. This study utilised co-occurrence networks in research categories and in keyword alongside with three visualisation views, namely Cluster View, which represent the distribution of research fields from diverse viewpoints, and Time-Line View and Time-Zone View, which illustrate the temporal evolution and interrelationships of research areas. The methodology ensured validity and reliability of the measurements, consistent with the approach.

4. Results and Findings

4.1. Time Series Segments of Publication Statistics

The volume of publications serves as a pivotal benchmark for discerning a field's developmental trajectory and prognosticating future directions (Figure 2).

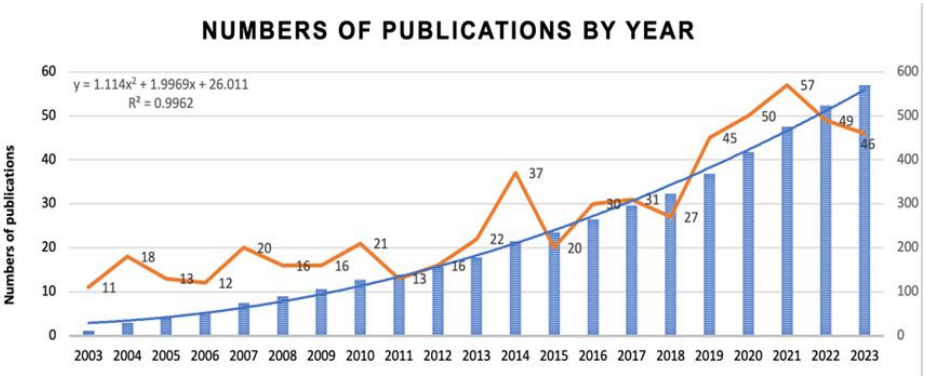


Figure 2. Publication statistics in time series segments (Authors' Own Source)

The orange line chart represents the annual incremental volume, while the blue bar chart denotes the cumulative amount. The blue exponential curve illustrates the trendline fitted through regression analysis. The PIM field has experienced growth, as the ascending trend in cumulative publications testifies. The growth trend aligns with the escalating interest from scholars in both interdisciplinary and cross-disciplinary studies. Despite this study only encompassing data from the initial seven months in 2023, a projection using linear regression estimates the total at 46 publications.

Overall, publications within the PIM domain show an upward trajectory, delineated into three phases: The Emerging Phase, Developing Phase, and Exploration Phase. During the Emerging Phase (2003-2012), the annual publication frequency showed variability, with an average of 16 publications per year. PIM, still in its nascent stage, attracted modest scholarly interest during this period. In the Developing Phase (2013-2018), there was a more robust publication output, with annual publications consistently exceeding 20 and peaking at 37 articles in 2014, reflecting a growing scholarly interest in cross-disciplinary research. The Exploration Phase (2019-2023) witnessed a pronounced surge in publications, averaging 49 articles per year. This surge underscores the increasing significance of PIM research, positioning it as a central area of academic inquiry and suggesting promising future growth.

4.2. Co-occurrence networks in research categories

By identifying cross-disciplinary and inter-disciplinary subjects within the PIM field, and observing their dynamic progression, It provides valuable guidance for future researchers exploring new directions.

Figure 3 visualises co-occurrence networks. Each node in the figure represents a category within the PIM domain, with larger nodes indicating higher occurrence. Thicker lines denote increased frequency of interdisciplinary research. Notably, the visualisation reveals Health Care as an isolated entity within the subject network, lacking intersection with other research categories, suggesting limited disciplinary crossovers.

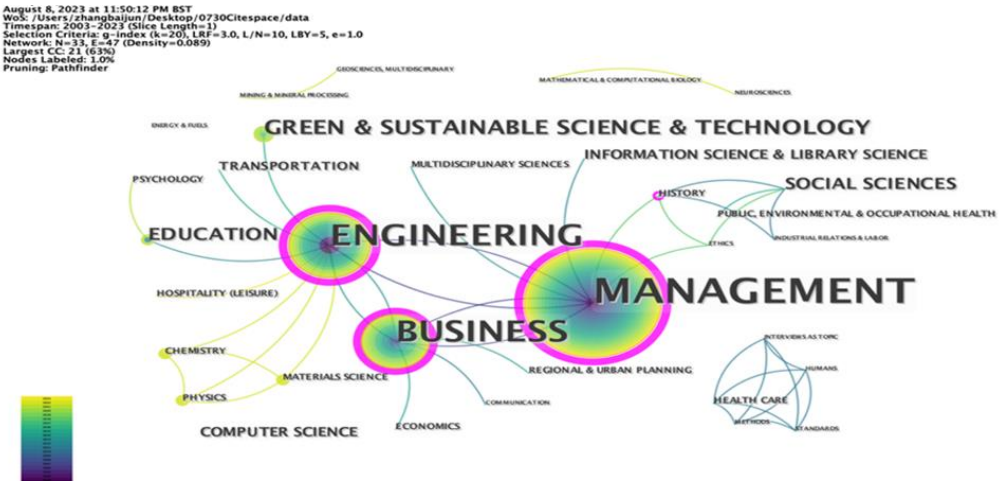


Figure 3. Co-occurrence networks in research categories (Authors' Own Source)

Disciplinary analysis unveils core and intersecting disciplines within the PIM domain, guiding future research directions. Figure 3 delineates key indicators for the top ten subjects in PIM publications, with Management, Engineering, and Business dominating but interdisciplinary collaborations remaining sporadic, suggesting modest disciplinary diversity. Conversely, fields like Healthcare, Neuroscience, Biotechnology, Geoscience, and Computer Science exhibit independence from the core disciplines, hinting at potential for diverse collaborations in PIM beyond conventional areas.

A notable prominence is observed on the "History" node, marked by robust centrality and a purple spotlight, primarily due to studies examining innovation initiatives through cultural and historical lenses. Responsible innovation necessitates consideration of broader socio-ethical and socio-economic implications (Flipse and van de Loo, 2018), indicating that future PIM research may continue converging at the intersections of history.

5. Keyword co-occurrence networks

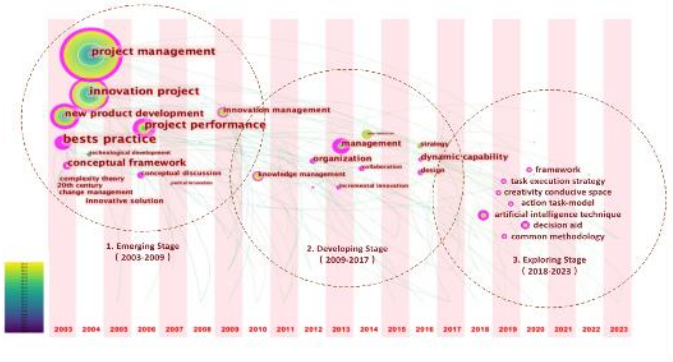
In the context of PIM, mapping keyword frequencies alongside their chronological occurrences reveals evolving trends in the field. This study utilised keyword clustering to pinpoint core research areas and assessed "burst" keywords to identify emerging research frontiers.

5.1. Keyword co-occurrence analysis

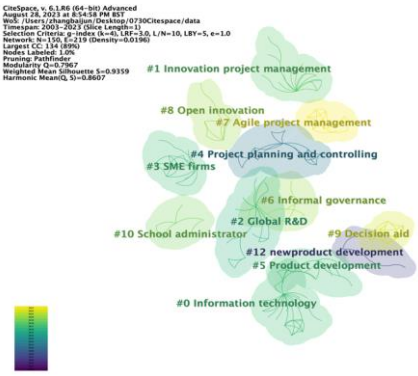
The Keyword co-occurrence network visualisation comprises 150 keywords and 219 links, suggesting robust keyword interactions can be seen in Figure 4.

Figure 4. Keyword co-occurrence networks (Authors' Own Source)

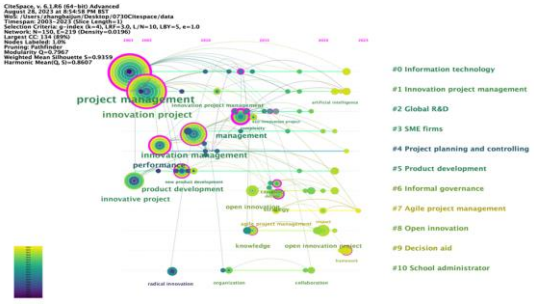
Analysis of the keyword frequency and centrality from Figure 4 reveals prevalent terms such as PM, IM, and innovation project. Significant nodes include new product development, project performance, knowledge management, and open innovation, indicating key focal points in PIM research. Hub nodes like best practice, conceptual framework, and dynamic capability serve as crucial connectors. Terms with a pink outer ring, like conceptual framework and incremental innovation, suggest future trends may emphasise framework establishment, enhanced management, and incremental innovation.


project management
innovation project
new product development
project performance
bests practice
conceptual framework
innovation management
management
organization
knowledge management
strategy
dynamic capability
framework
task execution strategy
creativity conducive space
action task-model
artificial intelligence technique
decision aid
common methodology
1. Emerging Stage
(2003-2009)
2. Developing Stage
(2009-2017)
3. Exploring Stage
(2018-2023)
2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023

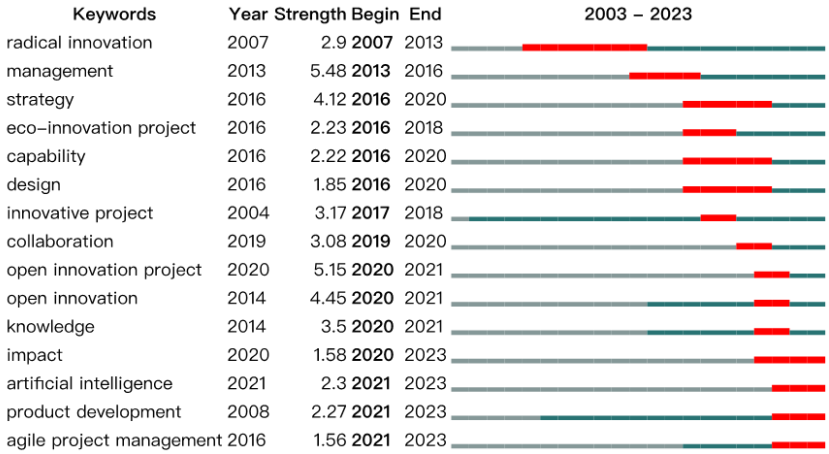
(a)


#1 Innovation project management
#8 Open innovation
#7 Agile project management
#4 Project planning and controlling
#3 SME firms
#6 Informal governance
#2 Global R&D
#10 School administrator
#9 Decision aid
#12 newproduct development
#5 Product development
#0 Information technology

(b)


project management
innovation project
management
innovation management
performance
product development
innovative project
#0 Information technology
#1 Innovation project management
#2 Global R&D
#3 SME firms
#4 Project planning and controlling
#5 Product development
#6 Informal governance
#7 Agile project management
#8 Open innovation
#9 Decision aid
#10 School administrator

(c)



(d)

Figure 5. Keyword time-zone map (a), Keywords clustering map (b), Keyword timeline graph (c), The list of top 15 Keywords with the strongest citation bursts (d) (Authors' Own Source)

5.2. Keyword time-zone map

In CiteSpace, the Keyword co-occurrence network facilitates the creation of a keyword time-zone map, plotting keywords across 1-year time segments (Figure 5.a), enabling the identification of each keyword's inception in the PIM literature. The node size represents the keyword's frequency, while the links visualise the progression of research. The map reveals three distinct phases: (1) Emerging Phase integrates PM with innovation projects, highlighting IM knowledge prominence. (2) Development Phase emphasises project interconnectedness within organisational processes, leading to strategies supporting organisational innovation. (3) Exploring Phase explores newer research areas.

5.3. Keyword Clustering

For the keyword clustering maps the Log Maximum Likelihood algorithm was employed (Figure 5.b). Eleven distinct research dimensions occurred with this analysis.

5.4. Keyword Timeline Graph

Leveraging the keyword clustering analysis, the keyword timeline graph illustrates the evolutionary trajectory of research orientations within this domain, providing a comprehensive visualisation of the progression and transformation of focal keywords. Aligning with prior clustering exploration, the timeline graph delineates dynamic shifts of various keywords under 11 predominant thematic clusters (Figure 5.c).

5.5. Keyword Burst

Burst words, characterised by their pronounced frequency fluctuations within specific time intervals, act as indicators of evolving subject trends. 15 keywords manifesting significant 'burst' characteristics was discerned, as depicted in Figure 5.d with corresponding red line segment. Mapping these burst keywords against the three temporal phases offers corroborative insights. (1) Emerging Phase exhibited a constrained breadth, highlighted solely by the burst term “Radical Innovation”. (2) Developing Phase marked an expansion in research volume and diversity, introducing burst terms such as "Management," "Strategy," and "Eco-Innovation". (3) Exploring Phase observed a steep incline in both burst word occurrences and publication metrics, this phase foregrounded concepts such as "Collaboration," "Knowledge," and "Agile Project Management" as pivotal research subjects.

6. Discussion and conclusion

This study conducted a thorough scientometric analysis on 521 literature pieces from renowned databases to comprehend the convergence in the PIM research domain over the last two decades. The findings offer insights for both industrial decision-making and academic research trajectories. Publication metrics reveal a rising trend in the cumulative number of papers within the PMI domain, particularly in the past five years, indicating increased scholarly interest and potential for research. Furthermore, disciplinary distribution identifies Management, Engineering, and Business as predominant subject areas in PIM, with potential intersections with diverse disciplines such as History, Healthcare, Neuroscience, Biology, Geoscience, and Computer Science. Research landmarks, including highly-cited papers and contributions from prominent researchers, provide insightful reviews. The progression of research is delineated into three stages: emerging, developing, and exploring, focusing on unique attributes of innovation projects, diverse facets of managing them, and probing cutting-edge research areas. Dominant research themes are divided into three domains focusing on managing uncertainty, investigating various efforts, and exploring applicability in complex situations. Frontier research fields gravitate towards AI, product development, and agile product management, emphasising the effective incorporation of AI into innovation endeavours, alignment of product development with disruptive innovation and digital transformation, and application of agile product management across industries.

This study has limitations concerning data source, scope, and methodology. Primary data were sourced from three databases. Incorporating data from additional databases, like Dimensions, could yield different results. Future research would benefit from such extended data sourcing. The research focused solely on peer-reviewed articles and reviews in English, potentially overlooking valuable insights from diverse publication types and languages. The scientometric mapping approach used bears inherent limitations, including citation bias and a time lag in data.

Combining scientometric review with traditional systematic review in future studies could mitigate these limitations.

References

Abbasi, A., Altmann, J. and Hossain, L. (2011) 'Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures.' *JOURNAL OF INFORMETRICS*. Amsterdam: Elsevier, 5(4) pp. 594–607.

Ansoff, H. (2007) *Strategic Management*. Springer.

APM (2019) *Association of Project Managers Body of Knowledge (APMBoK) Seventh edition*.

Ben Mahmoud-Jouini, S., Midler, C. and Silberzahn, P. (2016) 'Contributions of Design Thinking to Project Management in an Innovation Context.' *PROJECT MANAGEMENT JOURNAL*, 47(2) pp. 144–156.

Bibarsov, K. R., Khokholova, G. I. and Okladnikova, D. R. (2017) 'Conceptual basics and mechanism of innovation project management.' *European Research Studies Journal*, 20(2) pp. 224–235.

Burns, T. and Stalker, G. M. (1994) *The Management of Innovation*. Oxford, New York: Oxford University Press.

Chen, C., Hu, Z., Liu, S. and Tseng, H. (2012) 'Emerging trends in regenerative medicine: a scientometric analysis in CiteSpace.' *Expert Opinion on Biological Therapy*. Taylor & Francis, 12(5) pp. 593–608.

Davies, A. (2014) 'Innovation and Project Management.' *In* Dodgson, M., Gann, D. M., and Phillips, N. (eds) *The Oxford Handbook of Innovation Management*. Oxford University Press, p. 0.

De Rezende, L. B., Blackwell, P. and Pessanha Gonçalves, M. D. (2018) 'Research Focuses, Trends, and Major Findings on Project Complexity: A Bibliometric Network Analysis of 50 Years of Project Complexity Research.' *Project Management Journal*. SAGE Publications Inc, 49(1) pp. 42–56.

Dodgson, M. and Gann, D. (2011) 'Technological Innovation and Complex Systems in Cities.' *Journal of Urban Technology*. Routledge, 18(3) pp. 101–113.

Gault, F. (2013) *The Oslo Manual*. Gault, F. (ed.) *HANDBOOK OF INNOVATION INDICATORS AND MEASUREMENT*. Cheltenham: Edward Elgar Publishing Ltd (Elgar Original Reference), pp. 41–59.

Goldhar, J. D. (1994) 'Mastering the Dynamics of Innovation: How Companies Can Seize Opportunities in the Face of Technological Change.' *Sloan Management Review*. Massachusetts Institute of Technology, Cambridge, MA, 35(4) p. 97.

Guerra Betancourt, K., de Zayas Pérez, M. R. and González Guitián, M. V. (2013) 'Bibliometric analysis of publications related to innovation projects and their management in Scopus, 2001-2011.' *Revista Cubana de Informacion en Ciencias de la Salud*, 24(3) pp. 281–294.

Khalife, M. A., Dunay, A. and Illés, C. B. (2021) 'Bibliometric Analysis of Articles on Project Management Research.' *Periodica Polytechnica Social and Management Sciences*, 29(1) pp. 70–83.

Lenfle, S. and Soderlund, J. (2019) 'Large-Scale Innovative Projects as Temporary Trading Zones: Toward an Interlanguage Theory.' *ORGANISATION STUDIES*, 40(11) pp. 1713–1739.

Mentzer, M. (1987) 'Structure in Fives - Designing Effective Organisations - Mintzberg,h.' *ACADEMY OF MANAGEMENT REVIEW*. Briarcliff Manor: Acad Management, 12(2) pp. 395–401.

Morris, P. W. G. (2013) *Reconstructing Project Management*. John Wiley & Sons.

Oliveira Lucena, J. P., Lago Alves, T. da C. and de Medeiros Junior, J. V. (2019) 'Project Governance: a bibliometric analysis of 2014 to 2018.' *REVISTA DE GESTAO E PROJETOS*. Sao Paulo: Univ Nove Julho, 10(1) pp. 107–125.

Pinto, F. A., Frank, A. G. and Paula, I. C. de (2011) 'Definição de diretrises de gerenciamento de projetos empregando a análise de agrupamento: Um estudo exploratório.' *In Congresso Brasileiro de Gestão de Desenvolvimento de Produto (8.: 2011 set. 12-14: Porto Alegre, RS).[Anais][recurso eletrônico].[Porto Alegre, RS: Departamento de Engenharia de Produção e Transportes da UFRGS], 2011.*

Robbins, P. and O'Connor, G. C. (2023) 'The professionalization of innovation management: Evolution and implications.' *Journal of Product Innovation Management*, 40(5) pp. 593–609.

Russo, R. F. S. M., Sbragia, R. and Yu, A. S. O. (2017) 'Unknown unknowns in innovative projects: Early signs sensemaking.' *BAR - Brazilian Administration Review*, 14(3).

Shenhar, A. J. and Dvir, D. (2007) 'Reinventing project management: The diamond approach to successful growth and innovation.' *RESEARCH-TECHNOLOGY MANAGEMENT*. Arlington: Industrial Research Inst, Inc, 50(6) pp. 68–69.

Silva, E. and Gil, A. C. (2013) 'Inovação e Gestão de Projetos: Os "Fins" Justificam os "Meios."' *Gestão e Projetos: GeP*. Universidade Nove de Julho, 4(1). Gestão e Projetos: GeP pp. 138–164.

Silvius, G. (2017) 'Sustainability as a new school of thought in project management.' *Journal of Cleaner Production*, 166, November, pp. 1479–1493.

Utama, W., Chan, A., Zahoor, H. and Gao, R. (2020) 'Review of research trend in international construction projects: A bibliometric analysis.' *Construction Economics and Building*. UTS ePress, 16(2) pp. 71–82.

Wheelwright, S. and Clark, K. (1992) 'Creating Project Plans to Focus Product Development.' *HARVARD BUSINESS REVIEW*. Boulder: Harvard Business Review, 70(2) pp. 70–82.

Young, L., Ganguly, A. and Farr, J. V. (2012) 'Project management processes in agile project environment.' *In Annual International Conference of the American Society for Engineering Management*, pp. 9–19.

Zhang, L., Mohandes, S. R., Tong, J., Abadi, M., Banihashemi, S. and Deng, B. (2023) 'Sustainable Project Governance: Scientometric Analysis and Emerging Trends.' *Sustainability*. Multidisciplinary Digital Publishing Institute, 15(3) p. 2441.

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17691

Vaccine voices in the digital sphere: a multilayer network analysis of online forum discussion in Taiwan

Jason Dean-Chen Yin
School of Public Health, University of Hong Kong, Hong Kong, China.

How to cite: Yin, J. D-C. 2024. Vaccine voices in the digital sphere: a multilayer network analysis of online forum discussion in Taiwan. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17691

Abstract

New spatiotemporal considerations of the network allow surveillance of vaccine opinion online. The study uses a multilayer network – with each layer representing vaccine opinion discussion – to examine how the structure and timing of engagement in online communities may affect the spread of COVID-19 vaccine opinions. The aim is to improve public health messaging management during health crises and contribute to WHO's growing infodemic research agenda. The study finds that online discussions on COVID-19 vaccines are dominated by a few highly connected nodes within power-law structured communities. Vaccine-hesitant and pro-vaccine discussions are more engaged with, and more frequently posted earlier, but overall, less densely connected than the fewer, but highly clustered anti-vaccine discussions. Temporally, this trend increases over time for anti-vaccine discussions, suggesting insular communication (and potential echo chambering) happens gradually. The findings suggest proactive information management with consistent vaccine advocacy in online communities is crucial in low-activity periods, as dense anti-vaccination networks may pose misinformation risks.

Keywords: *multilayer network; vaccine hesitancy; infodemiology; social network analysis*

1. Introduction

The architecture and dynamics of network structure are influential in shaping the trajectory of information flow. These dynamics and structures have been explored in the digital realm in the age of the internet from various perspectives of network level. At the micro-network level, Bakshy et al. in their study of Facebook found that weak ties play a dominant role in disseminating information, highlighting the importance of users (a "node") in the network. At the meso-network level, the importance of community formation also has implications for information spread (Girvan & Newman, 2002). High degrees of clustering in communities can often form echo-chambers with limited exposure to other information, and facilitate stronger

information dissemination in those communities (Moody & White, 2003). At the macro-network level, the structure of node-edge connection describes the mechanisms for how the network grows. One such example is identifying if the network follows a power law distribution (Barabási, 2009). Those following power-law distributions are characterized by a few nodes having a disproportionate number of edges. Analyses of social media platforms have found that many follow power-law distributions (Mislove et al., 2007).

In the public health space, network studies have only recently become popular. For vaccine opinion in particular, studies have used network analysis to identify anti-vaccination themes (Featherstone et al., 2020; Lutkenhaus et al., 2019), study polarization in vaccine ideology (Jiang et al., 2021), or even track opinions on vaccines (Boucher et al., 2021; Gunaratne et al., 2019). While the COVID-19 pandemic has prompted more studies on networks in vaccine hesitancy, it is still a relatively understudied area.

One particularly overlooked area is in the spatiotemporal conceptualization of the network. Networks in vaccine hesitancy are usually represented as a single-layer network graph. However, a multilayer network offers a different lens to study information dissemination across different vaccine opinions. Conceptualizing the network as a multilayer would allow comparisons across structural differences between opinions and their dynamics. While relatively common in biology and physics, their use is limited in studying vaccine opinions (one such exception is their use in disease prediction (Fügenschuh & Fu, 2023; García et al., 2022).

Another overlooked area is incorporating the temporal dimension into network analysis. Networks are not static entities; rather, they change over time. This evolution can influence how information spreads across the network and shape the trajectory of dissemination. One relevant concept here is that of the "first-mover advantage" (Lieberman & Montgomery, 1988). In discussion on vaccines, first narratives may set the tone for ensuing conversations on vaccines. This timing may be important for managing misinformation, since early "psychological inoculation" can strengthen resistance against vaccine misinformation (Compton, 2013). Recent calls from vaccine hesitancy researchers to better incorporate elements of time into vaccine hesitancy research emphasize the salience of the temporal component (Larson et al., 2022).

The current study addresses the following questions: what are the patterns, observations, and implications of online community formation and their timing of online engagement on the spread of COVID-19 vaccine opinions? Specifically, it addresses the following two sub questions: 1) What are implications of differences in the structure and cohesion of different layers on vaccine opinion transmission; and 2) What are the implications of differences in timing on discussion engagement between the different opinions on information transmission? The findings provide insight on the targeted management and monitoring of online public health messaging for current and future public health crises.

2. Methods

2.1. Data

PTT is a terminal-based bulletin board system in Taiwan. It is a free and open forum and non-commercial, where users post and discuss a variety of topics. Often termed Taiwan's "Reddit", it is one of the most active forums in Taiwan. From July 2022 to July 2023, the average users per day was 56,000 (PTT, 2023).

The web-based version of PTT has the following structure. In the forum, there are *boards* and *board masters*, which are the same as subforums and moderators. Posts within each board can be done by creating a new post or a reply post. Within each post, users may leave follow-up comments in text that are also embedded with a sentiment: they can *like, boo,* or have a *neutral reaction*. These are like upvotes or likes, downvotes, or neutral replies, respectively.

All posts from the "Gossiping" board – the most active and popular on PTT – are collated from January 1, 2021 to December 1, 2022: dates that capture the vaccine stockpiling and administration during COVID-19. To find vaccine-related boards, the filter word *"vaccine"* in Chinese is used. For each discussion board, two independent labelers assigned a label of either "pro-vaccine", "vaccine hesitant", or "anti-vaccination" to delineate the sentiment of the post. These labels were classified using a combination of two main criteria. First is WHO's scope of "vaccine hesitant" to encapsulate a broad spectrum of reasons for non-vaccination (Larson et al., 2022; MacDonald et al., 2015). Second is consulting health psychology theories like Health Belief Model (Champion & Skinner, 2008) that include inaction due to self-efficacy reasons. Irrelevant posts mentioning vaccines but not related to vaccine discussions were discarded. The goal was a target inter-rater agreement of 85% and above, with discrepancies resolved by the main author. These three layers constitute the three vaccine opinions of the multilayer network. To construct the network, the embedded sentiment is used to assign the connection to a given layer. For those who "like" or leave neutral comments, they are assigned into the same layer as the sentiment of the post. For those who "boo", they will randomly be assigned into one of the two other layers. Given this randomness, the sample will be bootstrapped, and all measurements (elaborated below) calculated across bootstrap samples.

2.2. Data analysis

The multilayer M consists of three layers $M = \{G^P, G^H, G^A\}$, each being a directed, weighted graph that represents the aggregate links between commenters and authors for vaccine stances *pro-vaccination*, *vaccine hesitant*, and *anti-vaccination.* Each layer $l \in M$ consists of all interactions of the set of nodes V^l and set of edges E^l, and is represented as $G^l = (V^l, E^l)$, with each node being a user, and each edge being a comment on that user's post. These three layers represent the sentiments towards vaccination.

Mathematically, for the evolving network, assume that we observe the network over a finite time T, with starting point $t_s = 0$ and ending point $t_e = T$. Each layer in M is defined as $G^l_{0,T} = (V, E_{0,T})$ on a time interval $[0, T]$ which consists of a set of nodes or vertices V and a set of temporal edges $E_{0,T}$. The evolving multilayer network is thus $M_{0,T} = \{G^P_{0,T}, G^H_{0,T}, G^A_{0,T}\}$. This multilayer, temporal network is observed at discrete time points $t_1 t_2, \dots, t_{n-1}, t_n$. At any time point t_n, an instantiation of the multilayer, M_n, is observed, whereby each G^l_n contains the set of temporal edges E^l_n such that $(u, v)^l{}_{t_n} \in E^l_{0,T}$ with edges between nodes u, v contained within the period $t_n = [\, t_{n_s}, t_{n_e}]$ such that $t_{n_s} \leq T$ and $t_{n_e} \geq t_{n_s} \geq 0$ (*i.e.* the instantiation time is between the start time t_s and end time t_e, and the end time t_e is later than the start t_s). The graphs, being directed, are also non-mirrored such that $(u, v) \neq (v, u)$). Each of the following measures below are written for one snapshot t_n, but are calculated temporally.

Degree (average)

The degree of a node is the number of edges connected to it. The average degree of the layer is the average of degrees of all nodes in the network layer. For a directed graph $G(V, E)$, the average degrees $\bar{d}_{tot}$ of all nodes is:

$$\bar{d}_{tot} = \frac{\bar{d}_{in} + \bar{d}_{out}}{2} \tag{1}$$

where $\bar{d}_{in}$ and $\bar{d}_{out}$ represent the average indegrees and outdegrees of a given network.

Density

The density of a network layer measures the connectedness of the graph. For any network, the density $d(G)$ of a graph $G(V, E)$ is the number of edges divided by the theoretical possible number of edges. In a directed graph, the density is calculated as $\frac{E}{2V(V-1)}$, where E is the number of edges and V is the number of vertices (nodes). The density of a network ranges from 0 to 1, with 0 representing no edges, and 1 representing all possible edges present.

Power law distribution

A network is considered a power law network when the probability distribution of degree d, $p(d)$, follows a power law $p(d) \propto d^{\delta}$, where δ is the exponential parameter of the power law distribution, usually falling between $1 \leq \delta \leq 3$. Delta values around 2 indicate a power law network. This will be tested with the "powerlaw" library in Python.

Modularity

Modularity is defined as the difference between the actual number of edges within communities from the expected number of edges given the *degree distribution*, or, the number of edges connected to a node. Mathematically, modularity Q is defined as:

$$Q = \frac{1}{2m}\sum_i \sum_j \left[\frac{A_{ij} - k_i^{out} * k_j^{in}}{2m}\right] * \Delta(c_i, c_j) \quad (2)$$

where A_{ij} is the weight of the edge from node i to node k, k_i^{out} is the sum of weights of the edges leaving node i , k_j^{in} is the sum of the weights entering node j, m is the total weight of all edges in the graph, and $\Delta(c_i, c_j)$ is the Kronecker delta function, equalling 1 if nodes i and j belong to the same community, and 0 if not (boolean operator for contributing to modularity score).

Active communities

An "active community" is one that has a significant number of nodes compared to other communities in the same layer. If n^l and c^l denote the number of nodes and communities in layer l, and n_c^l represents the number of nodes in community c of layer l, then the existence of an active community (AC) in layer l is given as Boolean classifier:

$$AC^l(c) = \begin{cases} true, & if\ n_c^l \geq (\frac{n^l}{c^l}) \\ false, & else \end{cases} \quad (3)$$

Percent constitution of network layer by active community

After identifying the active communities, the number of nodes in the active communities is divided by all the nodes in the layer to calculate the percent constitution of network layer (PC_{AC}^l) by active communities with the following formula:

$$PC_{AC}^l = \frac{\sum_{c \in AC^l} n_c^l}{n^l} \quad (4)$$

Survival analysis

To analyse engagement, I fit the time-to-posting of each layer to a Cox Proportional Hazards Model, with predictor as layer. Hazard ratios of each possible pairwise comparison indicate the relative rate at which the groups experience the two events above, assuming proportional hazards.

3. Results

I find that all three layers have delta values – the power law parameter –significantly around two, with p-values above 0.05 ($G^P: \delta = 2.2, p = 0.067$ $G^H: \delta = 2.3, p = 0.068$ $G^A: \delta = 2.3, p = 0.09$). This confirms these layers are power-law networks, and all three layers generally revolve around a few highly connected nodes.

Discussions on the vaccine hesitant are the most active, followed by the pro-vaccine layer (Table 1). The layer with the least users and interactions is anti-vaccination. This is also reflected in

average degrees. These findings suggest more hesitant and positive discussions are more engaged with than anti-vaccination ones. However, density is in the opposite direction. The anti-vaccination layer has the highest density suggesting that, despite its small size, the discussions are more tightly connected as a network. Since degree and density are network size-dependent, the results in Table 1 should be interpreted cautiously. When comparing just the pro-vaccination and hesitant layers, which are more similar in size, the hesitant layer has more connections per node (higher degree) but is overall more dispersed (lower density), suggesting a more scattered network of information hubs.

Table 1. Properties of the multilayer network by layer (bootstrap n=1,000)

Layer	Number of posts	Nodes (average)	Edges (average)	Density (10^{-4})	Average degree	Modularity	Active communities	Active communities' constitution of network
G^P	1,283	11,087	23,504	1.91 (1.89 – 1.94)	4.24 (4.21 – 4.27)	0.574 (0.570 – 0.578)	46.0 (39.0 – 53.0)	79.6 (72.7 – 86.7)
G^H	1,322	14,037	33,520	1.70 (1.68 – 1.72)	4.77 (4.75 – 4.80)	0.543 (0.538 – 0.548)	34.0 (28.0 – 41.0)	75.9 (69.7 – 81.3)
G^A	387	5,477	8,696	2.90 (2.85 – 2.95)	3.18 (3.15 – 3.20)	0.699 (0.691 – 0.704)	21.0 (17.0 – 25.0)	57.4 (51.2 – 63.4)

On modularity, the pro-vaccination and vaccination communities are similar, while the anti-vaccination network is higher, indicating stronger clustering. Interpreted in tandem to density, the high density and modularity of the anti-vaccination layer suggests many localized and highly intra-connected communities that communicate amongst themselves. The pro and hesitant communities, alternatively, connect more sparsely in the communities but more strongly in the layer. The difference in the number of communities and their percent constitution of the overall network further illustrates this. The percent constitution of the entire layer is much lower for the anti-vaccination layer, suggesting that the active communities are dominant and main contributors to high density, while the other half of the layer is very inactive and separate from these tight communities.

Temporally, Figure 1 illustrates that anti-vaccination community cohesion grows over time, indicating insular information transmission hubs. On the other hand, the pro-vaccination and hesitant layers have steadier community cohesion over time. The gradual tapering in all three

layers in modularity and percentage suggests that new participants are isolated from these conversations, with this more noticeable in the anti-vaccination network.

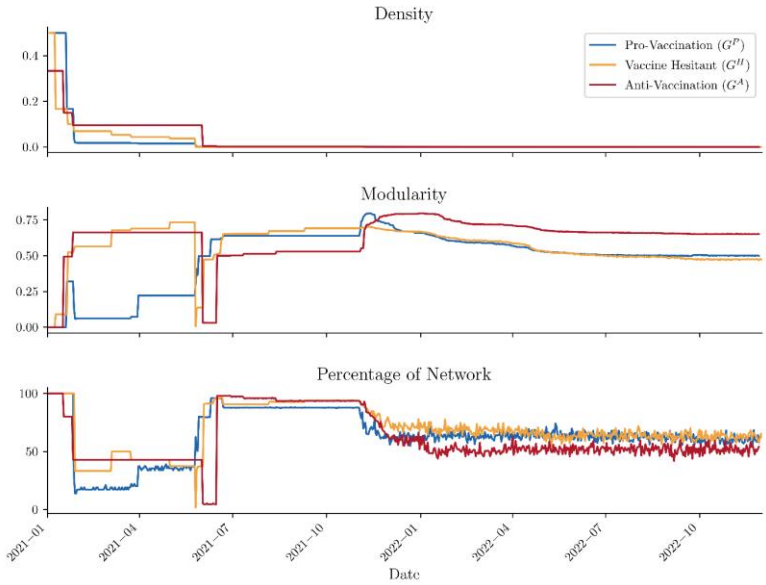


Figure 1. Temporal changes in density, modularity, and percentage of network occupied by active communities

Differences between topic starting were tested for each layer using survival curves and a Cox proportional hazards model to test earliness to conversation. Overall, the hesitant layer and anti-vaccination layers posted less than the pro-vaccination layer (0.87, p=0.005; 0.75, p=0.01, respectively, Table 2). Hazards were proportional in each layer. However, the discussion around vaccines only escalated much later around September 2021 after vaccines were procured, and the vaccination campaign was in effect on the island (Figure 2). In the period prior, the forum paid little attention to COVID-19 vaccinations.

Table 2. Hazard ratios for time-to-thread posting for each layer

Measure	Hazard ratio	p-value	Coefficient	Standard error	Z
Pro-vaccination (baseline)	-	-	-	-	-
Vaccine hesitant	0.87 (0.78 – 0.96)	0.005	-0.144	0.05	-2.79
Anti vaccination	0.75 (0.60 – 0.94)	0.014	-0.27	0.11	-2.41

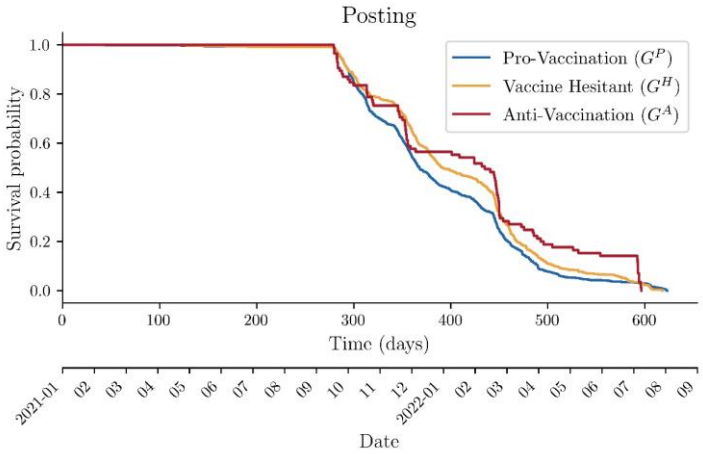


Figure 2. Thread-posting survival curves for the three layers

4. Discussion

In this paper, I have proposed and explored using a multilayer network to describe the formation of discussions online. I found that the anti-vaccination discussions are fewer, but more densely interconnected, indicating a higher risk of anti-vaccination misinformation in this group. This trend was only later in the time series, and not true initially. In addition, vaccine-hesitancy posts are made less often, but more sporadically. These findings underscore the importance of consistent vaccination advocacy in 'downtime' periods. This advocacy can target tightly knit anti-vaccination discussion communities that potentially form echo chambers. Findings in this study should be compared – both within and across platforms domestically, and internationally – to better triangulate findings.

References

Barabási, A. L. (2009). Scale-free networks: A decade and beyond. Science, 325(5939), 412–413.

Boucher, J.-C., Cornelson, K., Benham, J. L., Fullerton, M. M., Tang, T., Constantinescu, C., Mourali, M., Oxoby, R. J., Marshall, D. A., Hemmati, H., Badami, A., Hu, J., & Lang, R. (2021). Analyzing Social Media to Explore the Attitudes and Behaviors Following the Announcement of Successful COVID-19 Vaccine Trials: Infodemiology Study. JMIR Infodemiology, 1(1), e28800. https://pubmed.ncbi.nlm.nih.gov/34447924/

Champion, V. L., & Skinner, C. S. (2008). The Health Belief Model. In Health Behavior and Health Education: Theory, Research, and Practice (4th ed.). Jossey-Bass.

Compton, J. (2013). Inoculation theory. In The SAGE handbook of persuasion: Developments in theory and practice (pp. 220–236). Sage Publications, Inc. https://psycnet.apa.org/record/2013-39243-014

Featherstone, J. D., Ruiz, J. B., Barnett, G. A., & Millam, B. J. (2020). Exploring childhood vaccination themes and public opinions on Twitter: A semantic network analysis. Telematics and Informatics, 54, 101474.

Fügenschuh, M., & Fu, F. (2023). Overcoming Vaccine Hesitancy by Multiplex Social Network Targeting. Studies in Computational Intelligence, 1077 SCI, 576–587. https://link.springer.com/chapter/10.1007/978-3-031-21127-0_47

García, Y. E., Mery, G., Vásquez, P., Calvo, J. G., Barboza, L. A., Rivas, T., & Sanchez, F. (2022). Projecting the impact of Covid-19 variants and vaccination strategies in disease transmission using a multilayer network model in Costa Rica. Scientific Reports 2022 12:1, 12(1), 1–9. https://www.nature.com/articles/s41598-022-06236-1

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America, 99(12), 7821–7826. https://www.pnas.org/doi/abs/10.1073/pnas.122653799

Gunaratne, K., Coomes, E. A., & Haghbayan, H. (2019). Temporal trends in anti-vaccine discourse on Twitter. Vaccine, 37(35), 4867–4871. https://pubmed.ncbi.nlm.nih.gov/31300292/

Jiang, X., Su, M. H., Hwang, J., Lian, R., Brauer, M., Kim, S., & Shah, D. (2021). Polarization Over Vaccination: Ideological Differences in Twitter Expression About COVID-19 Vaccine Favorability and Specific Hesitancy Concerns: Social Media and Society, 7(3). https://journals.sagepub.com/doi/full/10.1177/20563051211048413

Larson, H. J., Gakidou, E., & Murray, C. J. L. (2022). The Vaccine-Hesitant Moment. New England Journal of Medicine, 387(1), 58–65. https://www.nejm.org/doi/full/10.1056/nejmra2106441

Lieberman, M. B., & Montgomery, D. B. (1988). First-mover advantages. Strategic Management Journal, 9(S1), 41–58. https://onlinelibrary.wiley.com/doi/full/10.1002/smj.4250090706

Lutkenhaus, R. O., Jansz, J., & Bouman, M. P. A. (2019). Mapping the Dutch vaccination debate on Twitter: Identifying communities, narratives, and interactions. Vaccine: X, 1, 100019.

MacDonald, N. E., Eskola, J., Liang, X., Chaudhuri, M., Dube, E., Gellin, B., Goldstein, S., Larson, H., Manzo, M. L., Reingold, A., Tshering, K., Zhou, Y., Duclos, P., Guirguis, S., Hickler, B., & Schuster, M. (2015). Vaccine hesitancy: Definition, scope and determinants. Vaccine, 33(34), 4161–4164. https://pubmed.ncbi.nlm.nih.gov/25896383/

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2007). Measurement and analysis of online social networks. Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC, 29–42. https://dl.acm.org/doi/10.1145/1298306.1298311

Moody, J., & White, D. R. (2003). Structural cohesion and embeddedness: A hierarchical concept of social groups. American Sociological Review, 68(1), 103–127.

PTT. (2023). PTT Online User Statistics. https://www.ptt.cc/bbs/Record/search?q=上站人次統計

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17765

The Invasion of Ukraine Viewed through Large-Scale Analysis of TikTok

Benjamin Steel [1], Sara Parker [2], Derek Ruths [1]

[1]Department of Computer Science, McGill University, Canada, [2]Media Ecosystem Observatory, McGill University, Canada.

How to cite: Steel, B.; Parker, S.; Ruths, D. 2024. The Invasion of Ukraine Viewed through Large-Scale Analysis of TikTok. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17765

Abstract

The vast majority of TikTok analysis done to date has used small, manually-collected datasets. This leaves open the questions of how to do large-scale analysis of phenomena on TikTok and what can be learned about unfolding current events. In this paper, we seek to better understand how such events present themselves on TikTok by conducting a first examination of large-scale user and content dynamics around the invasion of Ukraine in 2022. As this is among the first studies to conduct large-scale (i.e., involving millions of data points) data collection and analysis of TikTok data, our contributions also include insights into best (and less-than-great) practices for such critical research tasks. Furthermore, we have open-sourced our Ukraine-invasion dataset and provide a software library that can be used to collect TikTok data.

Keywords: *tiktok; social media; Ukraine.*

1. Introduction

TikTok has rapidly become socially, politically, and economically important. The platform now boasts over one billion active users, and over a quarter of people below age 25 in the United States consider TikTok their primary news source (Matsa, 2022, Stokel-Walker, 2022). As has been witnessed on other social platforms, at this scale of adoption, trends on TikTok can have society-scale effects. There is urgency, then, in answering key questions: how can large-scale analysis be done on such a platform to study society-scale dynamics? What can be gleaned from such analysis about human behavior surrounding current events?

These questions are largely unaddressed. Existing research on TikTok has almost exclusively focused on small, curated datasets. This is due, in part, to the challenges posed by the design and novelty of the platform as compared to Twitter, Reddit, and other microblog-centric platforms. We considered these questions by undertaking the first large-scale study of TikTok content dynamics around the evolving Russian invasion of Ukraine - ultimately building and

working with a dataset of 9,500 (video) posts, 4.4 million comments, and 2.6 million users. This study forced us to engage with a number of methodological challenges that have long been settled on other platforms like Twitter - specifically, how to collect a representative dataset. As a result, part of our work provides an early guide for researchers seeking to use TikTok as a medium or object of study. With our dataset in hand, we looked at how language usage changed during the first months of the invasion, the dynamics of invasion-related topics, and capability of existing bot detection methods on the platform.

We make four contributions. First, we establish that TikTok manifests deeply contentious aspects of geopolitical events. Second, our findings make clear the need for new bot detection systems and paradigms for establishing dataset representativeness for TikTok. Third, this study provides a (albeit imperfect) template for TikTok data collection. And fourth, this study provides a dataset and data collection library: github.com/networkdynamics/pytok for future studies using TikTok.

2. Background

Prior studies of TikTok relating to the invasion of Ukraine have tended to involve small, manually collected datasets, typically involving between 50 and 500 posts, often without consideration of comments or metadata for those videos (ElHawary, 2023; Primig et al., 2023; Badola, 2023). Of existing TikTok studies, the largest we know of is (Medina Serrano et al., 2020) which studied partisan behavior using 8000 posts. To the best of our knowledge, this is the largest dataset of content on TikTok relating to the invasion of Ukraine by orders of magnitude. Our intent here has been to leverage this scale in order to assess phenomena that would be difficult or impossible to reliably measure with smaller datasets.

3. Dataset

The preparation of our dataset involved three steps: (1) building software for collecting data from TikTok, (2) designing and executing a methodology for collecting a broad, topically-coherent set of posts, and (3) filtering the data obtained. Work still needs to be done to improve TikTok data collection - we offer our process as a step along that path. We have released the data and code required to rebuild the dataset here: github.com/networkdynamics/ukraine-tiktok.

3.1. Collection Software Library

We required a method that both searches and downloads TikTok content. While TikTok does have a public API, it is not yet widely accessible. We investigated TikTok scraping libraries but found that none fit our needs: some libraries are browser automation-based, which results in slow collection times, while others are entirely requests-based, which is vulnerable to TikTok

backend API changes. We therefore developed an open-source library github.com/networkdynamics/pytok based off of github.com/davidteather/TikTok-Api. Our approach strikes a balance between prior approaches: browser automation for initial access, HTTP requests for fast access, and fallbacks to browser automation again if this fails. Our library has already been used by multiple other researchers, showing its demand.

3.2. Collection Methodology

Data access methods are limited on TikTok compared to Twitter, Reddit, or other platforms, with the main pathways being algorithmic feed, user-based search, and keyword-based search. Due to the black-box nature of the TikTok recommendation algorithm, we opted for a keyword-based search method. Using our library, we collected videos using a combination of hashtag and general search functionality. Hashtag searches are limited to the 1000 most viewed videos with that hashtag, so to expand the video set, we also used the general search functionality, which generally returns more videos for a search term. This approach is still non-ideal for data collection: the fuzzy nature of the black-box search functionality can return content unrelated to the search term; additionally videos appeared to be ordered in no available parameter-based order. Despite these limitations, we considered this the best of available options. We used a *seed-and-snowball* approach to expand the list of search terms: we did an initial informal search of TikTok, to find the most common hashtags used in popular videos related to the war, and we used this set as the seed set of hashtags (the first ten in the list below). We collected videos tagged with these terms, and examined the ranked co-occurring hashtags in the collected video descriptions to expand our search terms. The final hashtag set (with translations, and the corresponding language) was as follows: *standwithukraine*, *russia*, *nato*, *putin*, *moscow*, *zelenskyy*, *stopwar*, *stopthewar*, *ukrainewar*, *ww3*, володимирзеленський (Volodymyr Zelenskyy, ukr) , славаукраїні (Glory to Ukraine, ukr), путінхуйло (Fuck Putin, ukr), россия (Russia, ukr), війнавукраїні (War in Ukraine, ukr), зеленський (Zelenskyy, ukr), нівійні (No war, ukr), війна (War, ukr), нетвойне (No war, rus), зеленский (Zelenskyy, rus), путинхуйло (Fuck Putin, rus), *#denazification, #specialmilitaryoperation, #africansinukraine, #putinspeech, #whatshappeninginukraine*.

We ran this collection process in July 2022 and April 2023. The behavior of historical search in TikTok is currently uncharacterized - raising concerns over completeness. However, as we see in the count of content over time in Fig. 1, the number of results returned across our collection period was stable, implying that query timing does not greatly affect the sample. We also collected the comments for each of these videos (limiting to 1000 comments per video to ensure reasonable collection duration), to provide additional text data for analysis.

Due to the noise of the search functionalities, and hashtag abuse, there were some irrelevant videos in the raw dataset. We therefore removed this content from the data to produce the final

dataset, ensuring it contains only videos related to the invasion of Ukraine. We define "related to the invasion" as any video containing one of the following items:

Depictions or discussions of combat; Support or protest for either side's war efforts, including propagandistic content; Any mention of Putin or Zelenskyy during the invasion; Critical political and military leaders engaging with the invasion; Videos about direct social or economic outcomes of the war; Speculation about the war; Videos about the militaries of countries involved in the war posted during the invasion (as implied propaganda).

With this definition, we took a sample of 300 videos from the dataset, opened each video in TikTok, and manually labelled them as related or not. We found that of the videos that were still available, 63% were related to the war. 29% were no longer available. That such a high percentage of the content collected was not available 6 months after collection shows how ephemeral content is on TikTok. We fine-tuned a RoBERTa large language model (LLM) (Liu et al., 2019) using our 300 labelled videos descriptions to classify the videos as being related to the invasion or not, then used it to filter the dataset to only war related videos. For this filtered set, we found that of those still available, 93% were related to the war. Post-filtering, the full dataset contains text and metadata for approximately 9.5 thousand videos related to the invasion of Ukraine with 4.4 million comments, from 2.6 million users. Of users who posted a video, the mean number of videos posted was 1.7, with a max of 152. Of users who posted a comment, the mean number of comments was 1.7, with a max of 832. Given that we limited our max number of comments per video to 1000, the mean number of comments per video was 766. Fig. 1 shows the number of videos and comments in our dataset over time.

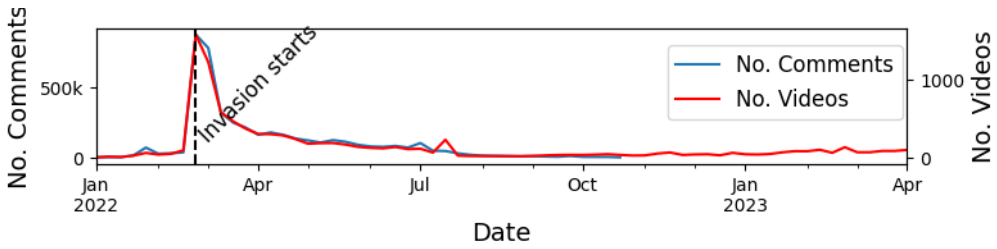


Figure 1: Number of videos and comments in our dataset over time.

4. Experiments

4.1. Words and Languages

We first took a macro view of how language evolved on the platform. Specifically, we wanted to measure the use of languages and words from the beginning of the invasion and onwards, to understand if there was any one changepoint at the start of the invasion, or if there were more complex, ongoing changes as communities and the platform itself responded to the unfolding

crisis. To do this temporal analysis, we used simple keyword, hashtag, and language searches in the comments we have from the videos, with language data provided by TikTok data. Where common keywords have multiple spellings, we have searched for all of these terms, and summed the counts to find the final search count. Note that these results are not meant to provide a comprehensive understanding of language patterns, but rather to show what possible new effects can be uncovered with large-scale analysis of TikTok data.

At this level of social interaction, we saw some behavioural changes over time. Notably, we saw evidence of mass movement to the Ukrainian language instead of Russian over time by users who at some point used Ukrainian, Fig. 2. We also saw a change from majority English text to Russian text over the course of the invasion in Fig. 3, indicating sustained attention from Russian speakers, or a decrease in attention from English speakers, or both. It seems likely that the medium of video allows greater mutual interaction between different language populations, allowing language dynamics that may not be seen on a text based platform.

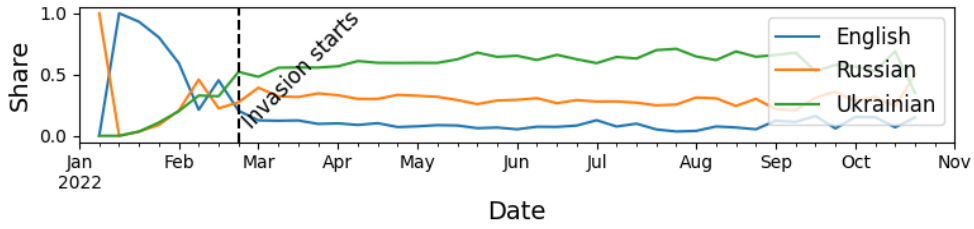


Figure 2: Language use for users who at some point use the Ukrainian language

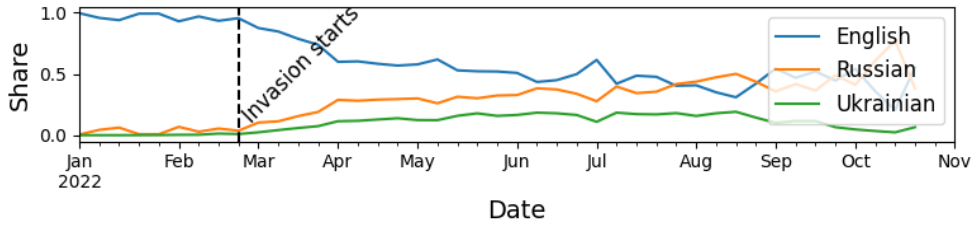


Figure 3: Language use over time

We also examined country and leader name mentions, finding that attention on Putin remained consistent throughout the war, but attention on Zelenskyy stays low (Fig. 4). Conversely, we saw sustained attention on Ukraine, but quickly diminishing attention on Russia in Fig. 5. This presents a curious juxtaposition among TikTok users: maintaining a focus on Ukraine and events within it while paying attention to Putin (rather than Zelenskyy).

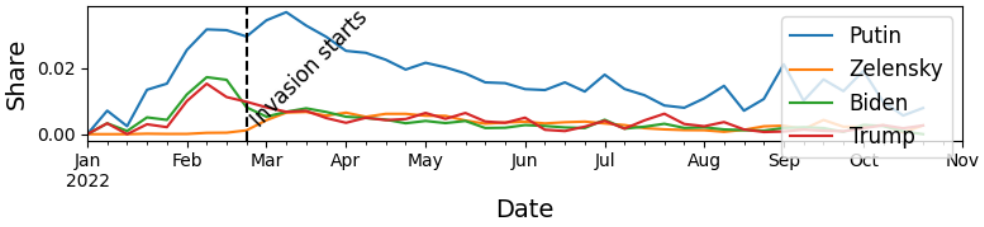


Figure 4: Leader mentions over time.

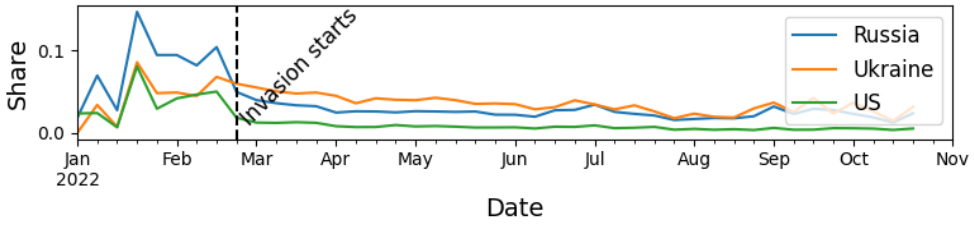


Figure 5: Country mentions over time.

4.2. Topics

We used topic modelling to examine video descriptions, as they provide us a view into the major themes on the platform. We used the BERTopic library (Grootendorst, 2022) and a multilingual Twitter fine-tuned LLM (DeLucia, 2022). In Fig. 6 we can see a diverse range of topics reflecting various political perspectives across the platform. Alongside popular topics aligned with general public media discourse that we would expect to see (NATO and Biden's relationship with the war, the position of Poland in the war), we also see more TikTok specific discussion, such as Eurovision's part in European solidarity and fears of an invasion of Alaska, indicative of nuanced discourse unique to the platform.

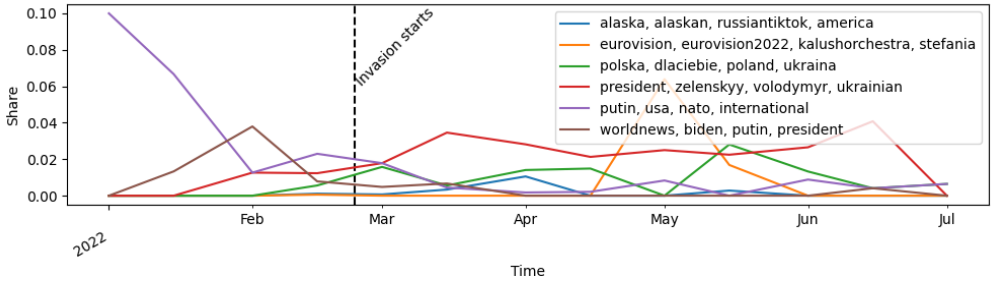


Figure 6: Selection of topics over time.

4.3. Bots

Bot detection is a crucial tool in large-scale social media analysis. We used a free and open source Twitter bot classifier focusing on generalizability (Ram, 2021), operationalizing the classifier features with the closest TikTok feature. We did not anticipate that this model would work perfectly on TikTok data, but we wanted to understand how poor the performance would be. In short, the performance was very bad. 99.2% were scored as likely to be a bot by the bot detection system, which, looking at the data, is unlikely. Examining features that were most attended to by the classifier, we find that verified status, following count, and account age are the top contributors to these classifications. It's clear that the predictive relationships between being a bot or not are very different between TikTok and Twitter.

5. Discussion

TikTok manifests intriguing social dynamics around current events. Where language is concerned, we found clear shifts in population use of languages that track and are explainable within the context of the invasion of Ukraine. That geopolitical events would manifest themselves in changes in language use is a striking example of the unexpected cultural impacts we can uncover through large-scale social media analysis. The large-scale dynamics in topical engagement shows how TikTok reflects discourse around the unfolding invasion. Moreover, the dynamics surrounding country and leader attention underscores the phenomena that can emerge and be studied within platforms with a short video-sharing mechanism.

TikTok requires revisiting fundamental aspects of large-scale analysis. Our work highlighted the need for substantial additional work in (at least) two areas. First, we as a research community lack the frameworks and baseline statistics for assessing and designing representative datasets from TikTok. Second, tools for identifying bot-generated content are completely lacking. Until we develop methods to address them, these issues will hamper large-scale studies of TikTok. In spite of the limitations highlighted, we consider the data preparation process, the data collection library, and the dataset to be valuable resources for the community as we collectively improve our capacity to conduct large-scale studies on TikTok. As evidenced by our findings, this effort is warranted: there is a great deal of social, political, and economic value to be gleaned from society-scale studies of the TikTok platform.

References

Badola, P. (2023). Russia and Ukraine: A Content Analysis of "The World's First TikTok War".

DeLucia, A., Wu, S., Mueller, A., Aguirre, C., Resnik, P., & Dredze, M. (2022, December). Bernice: A multilingual pre-trained encoder for Twitter. In Proceedings of the 2022 conference on empirical methods in natural language processing (pp. 6191-6205).

ElHawary, D. M. M. (2023). TikTok Battlefield: Comparative Analysis of English and Arabic Language Representations of The 2022 Russian Ukrainian Conflict On TikTok (Doctoral dissertation, The American University in Cairo (Egypt)).

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Matsa, K. E. (2022). More Americans are getting news on TikTok, bucking the trend on other social media sites. Pew Research Center, 21.

Medina Serrano, J. C., Papakyriakopoulos, O., & Hegelich, S. (2020, July). Dancing to the partisan beat: A first analysis of political communication on TikTok. In Proceedings of the 12th ACM Conference on Web Science (pp. 257-266).

Primig, F., Szabó, H. D., & Lacasa, P. (2023). Remixing war: An analysis of the reimagination of the Russian–Ukraine war on TikTok. Frontiers in Political Science, 5, 1085149.

Ram, R., Kong, Q., & Rizoiu, M. A. (2021, March). Birdspotter: A tool for analyzing and labeling twitter users. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining (pp. 918-921).

Stokel-Walker, C. (2022). TikTok wants longer videos-whether you like it or not. Wired. com. URL: https://www. wired. com/story/tiktok-wants-longer-videos-like-not/[accessed 2022-07-10].

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17768

TikTok vs. the Fourth Estate: Engagement With News on TikTok

Sara Parker[1], Benjamin Steel[2], Derek Ruths[2]

[1]Media Ecosystem Observatory, McGill University, Canada. [2]Department of Computer Science, McGill University, Canada.

How to cite: Parker, S.; Steel, B.; Ruths, D. 2024. TikTok vs. the Fourth Estate: Engagement with News on TikTok. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17768

Abstract

In addition to content from accounts the user follows, TikTok frequently emphasizes content with similar subject matter to videos the user has previously liked. As long as a user has indicated an interest in the topic (most likely by engaging with a related video), they may see a TikTok about it despite not following anyone who has ever interacted with it. Consequently, political networks and communities do not always emerge around prominent figures like politicians or professional content creators, but rather manifest as ephemeral trending topics. Our methodological approach to studying engagement with politics on TikTok therefore uses topic modeling to identify how and when TikTok users respond to news coverage about prominent current events. Specifically, we examine a large dataset of news articles, TikTok videos, and TikTok comments to uncover how TikTok discussion of the Russian-Ukrainian war temporally differs from coverage in mainstream news outlets. By examining points in time where the proportion of TikTok content about a specific topic within the war mirrors its discussion in the news – and more importantly, where it diverges – we are able to see how TikTok users include news events in their engagement with political issues. We find that the majority of TikTok videos about the Russian-Ukrainian war rarely feature prominent news stories, but rather focus on the users' personal experiences and perspectives. However, TikTok comments are more strongly engaged with specific events and 'hot-button' issues related to the war. The application of our methodology allows us to observe how major news events inform the creation and discussion of content on TikTok, the discrepancy between video descriptions/hashtags and video subject matter, and the importance of the comment section as a site for political conversation.

Keywords: *TikTok, news, Russian-Ukrainian war, topic modelling*

1. Introduction

In just a few short years, TikTok has become a major topic in global politics. Now with almost two billion monthly active users worldwide (Aslam, 2024), the platform has become the focus of substantial discussion, criticism, and research about its potential role as a mediator of information, particularly about major political events that may inspire significant amounts of disinformation or misleading content. TikTok is also especially popular among youth, increasing the importance of understanding how the app may be influencing the thoughts and views of upcoming generations of voters. However, the question of how to study engagement with politics and current events on TikTok remains elusive due to the distinct affordances of the platform.

In addition to content from accounts the user follows, TikTok frequently emphasizes content with similar subject matter to videos the user has previously liked (Boeker and Urman, 2022). As long as a user has indicated an interest in the topic, most likely by engaging with a related video or user, they may see a TikTok video about it despite not following anyone who has ever interacted with it. Consequently, political networks and communities do not always emerge around prominent figures like politicians or professional content creators, but rather manifest as ephemeral trending topics with which anyone can engage by simply posting an unedited video explaining their viewpoint from their bedroom and potentially gaining significantly more engagement than a video produced by a news organization. Discussion about politics on TikTok is therefore untethered from a discernible central node, making it much more complicated to observe how users may be interacting with or discussing a given political issue. However, that discussion is certainly happening: Literat and Kligler-Vilenchik (2021) have observed that the average proportion of political comments on TikTok is higher than on YouTube and much higher than on Instagram.

The methodological approach to studying political events on TikTok presented in this paper therefore seeks to tie these political dynamics to something better understood: news coverage. Specifically, we examine a large dataset of news articles, TikTok videos, and TikTok comments to uncover how TikTok discussion of the Russian-Ukrainian war differs from traditional news coverage. We chose the Russian-Ukrainian war as a case study of a major political topic because it is a prolonged multi-dimensional event, allowing us to explore political information on TikTok over time and in relation to a variety of specific events and discussion topics. By examining points in time where the proportion of TikTok content about a specific topic within the war mirrors its discussion in the news – and more importantly, where it diverges – we are able to see how TikTok users include news events in their engagement with political issues.

We find that TikTok videos about the Russian-Ukrainian war rarely feature prominent news stories, but rather focus on the users' personal experiences and perspectives. However, TikTok comments are more frequently engaged with specific events and key sub-topics related to the

war. The application of our methodology allows us to observe how major news events inform the creation and discussion of content on TikTok, the discrepancy between video descriptions/hashtags and video subject matter, and the importance of the comment section as a site for political conversation.

2. Methodology

2.1. Data collection

We use a dataset of 5700 news articles from a variety of international news organizations, 9500 TikTok video descriptions, and 4.4 million TikTok comments.

To find news articles related to the Russian-Ukrainian War, we searched for the keywords "ukraine", "russia", "putin", "zelensky", "kyiv", "moscow", and "nato" on AP News, The New York Times, The Washington Post, BBC, Reuters, Aljazeera, CNN, Fox News, Breitbart, MSNBC, The Wall Street Journal, Huffington Post, Vox, NBC News, USA Today, and NPR. We then collected 100 articles for each combination of keywords and news sites. We collected these articles using the seldonite library on GitHub, which uses the Google Custom Search JSON API for news searches on Google. The library also allows constraining of date range and site searches, which we used to filter the articles from January 2022 to April 2023, producing 5700 articles after de-duplication. A sanity check of the documents confirmed that the vast majority of articles were related to the invasion of Ukraine.

For our TikTok data, we used a dataset from Steel et al. (2023). This dataset contains videos and comments from January 2022 to April 2023, with videos collected via a hashtag-based search method, including the hashtags "#standwithukraine", "#russia", "#nato", "#putin", "#moscow", "#zelenskyy", "#stopwar", "#stopthewar", and "#ukrainewar". The videos were then filtered for relevance to the Russian-Ukrainian war using a fine-tuned language model trained on annotated data, resulting in 9500 videos, with 4.4 million corresponding comments. Due to data collection constraints imposed by TikTok, we could not analyze comments posted after October 2022.

2.2. Comparing topic proportion

We identified eighteen terms that related to key sub-topics of the war to examine how news coverage of major events during the war compare to discussion about those topics on TikTok: "ukraine", "russia", "putin", "zelensky", "kyiv", "moscow", "crimea", "donetsk", "donbas", "nato", "kharkiv", "odesa", "chechen", "azov", "kremlin", "denazification", "mariupol", and "kherson". Where keywords had multiple spellings, we searched for all those spellings, them summed the results (e.g., "odessa" and "odesa"; "kyiv" and "kiev"). Using keyword searches on our dataset, we identify the number of media (article, video, or comment) that mentioned a

keyword at least once. For the news articles, we searched both the title and text for mention of these keywords; for videos, we searched the video description; and for comments, we searched the comment text.

We then compare the proportion of each keyword per media type over time. We are less interested in how often topics were mentioned and more interested in when they were mentioned. The simultaneous increases and decreases in topic mentions across news, videos, and comments provides insight into how and when specific topics enter the TikTok discussion sphere. For example, simultaneous increases and decreases in topic mentions in news and videos would suggest that TikTok creators are likely quickly responding to news events when they create their content. Meanwhile a lack of concurrence in topic mentions between news and TikTok suggests that creators pay little attention to news content and are instead inspired by other things, like the potential for viewership and engagement, when creating and tagging videos. Furthermore, correlations between comments and news would suggest that TikTok users are in fact paying attention to news events, even if the videos themselves do not reflect news coverage. Additionally, a lack of concurrence between comments and videos would suggest that non-creator TikTok users use the TikTok comment section to engage in conversation and debate, even if the video itself is not explicitly related to the topic.

2.3 Analytical method

After identifying two-week segments where the news, TikTok video, and TikTok comment discussion about key topics significantly overlaps or diverges, we extract samples of videos and comments for the keywords and times of interest. We manually review hundreds of the most viewed videos and most liked comments from these samples to gain a contextualized understanding of the content and identify how TikTok users were actually engaging with the topics.

3. Results

First, we see that the number of TikTok comments and videos in the dataset increase and decrease at the same times as news content, showing that there is simultaneity between news coverage of a major event like the Russian-Ukrainian war and TikTok engagement with it (Fig. 1). However, we also see a much more prominent peak of TikTok content at the start of the war, with news content being more sustained. It is difficult to know the cause of the underlying change in content numbers compared to the effect that the TikTok search function has on amplifying content from certain times, but it may be that news content maintained more sustained coverage of the invasion, compared to the flurry of coverage from TikTok of the invasion in February and March of 2022. Regardless, the varying amount of content has implications for the significance of results at different points in the conflict.

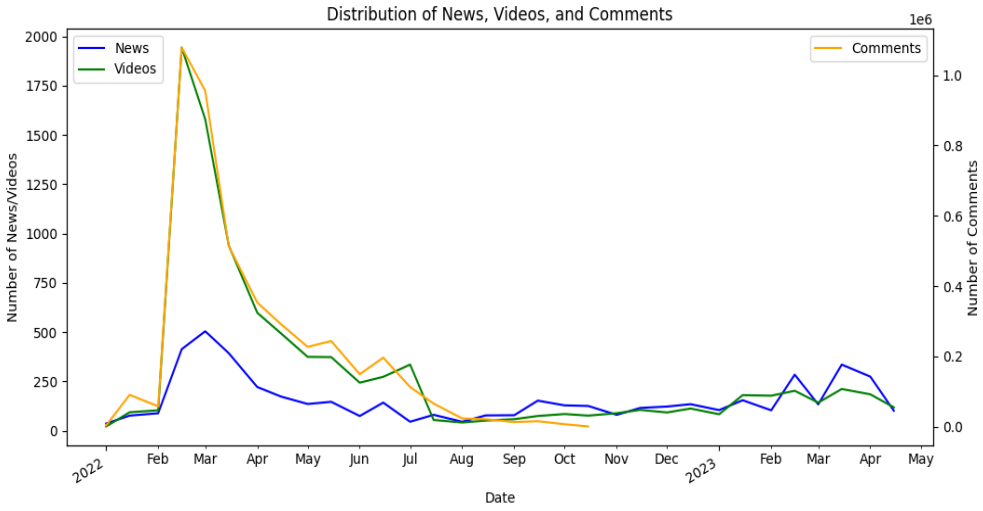


Figure 1: Distribution of news articles, TikTok videos, and TikTok comments in dataset from January 2022 to March 2023

3.1. Comparing videos and news

We see instances of simultaneous coverage on TikTok and in traditional news media for thirteen topics, suggesting that, at least part of the time, TikTok creators use news-relevant keywords in their video descriptions. Sometimes, the content is relevant to specific news events, particularly when using more specific hashtags. For example, videos posted in November 2022 with the keyword "kherson" appear to be responding to the news that the city had been liberated, with videos showing soldiers returning home and Zelensky's visit to the city (Fig. 2). However, TikTok content about a topic does not only increase with news content about that topic, suggesting that content creators are motivated by factors other than increased news coverage to publish content.

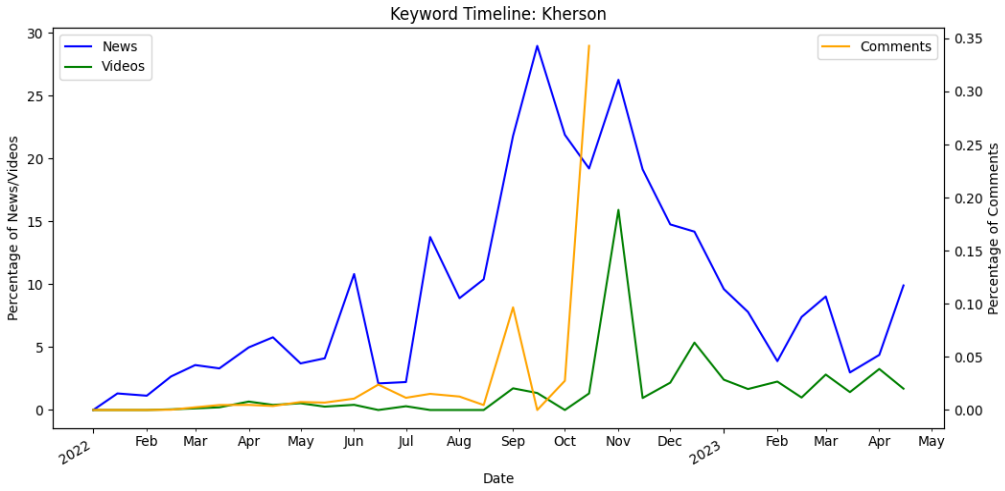


Figure 2: Timeline of mentions of "kherson" in News vs TikTok videos

Notably, many of the TikTok videos related to the Russian-Ukrainian war do not appear to be informative or specific to news events, but rather consist of slice-of-life content of people living or fighting in Ukraine, montages of Vladimir Putin or Volodymyr Zelenskyy, or videos of destruction across the country. For example, we find that, in March 2022, "putin" is mentioned in approximately 60% of news articles, over 20% of TikTok videos, and approximately 4% of TikTok comments. By examining how these proportions rise and fall over time, we can evaluate the importance of each topic in major news coverage, TikTok videos about the war, and comments. This comparison also allows us to identify how similar or distinct news coverage of a topic is to TikTok 'coverage' of the issues: for instance, although news coverage of the war appears quite focused on Russian President Vladimir Putin, TikTokers appear less interested in him (Fig. 3). Even when the proportion of videos using the keyword "putin" increases at times when news coverage of Putin increases, many of the TikTok videos do not appear to feature current footage of Putin or even explicitly mention a news event about him.

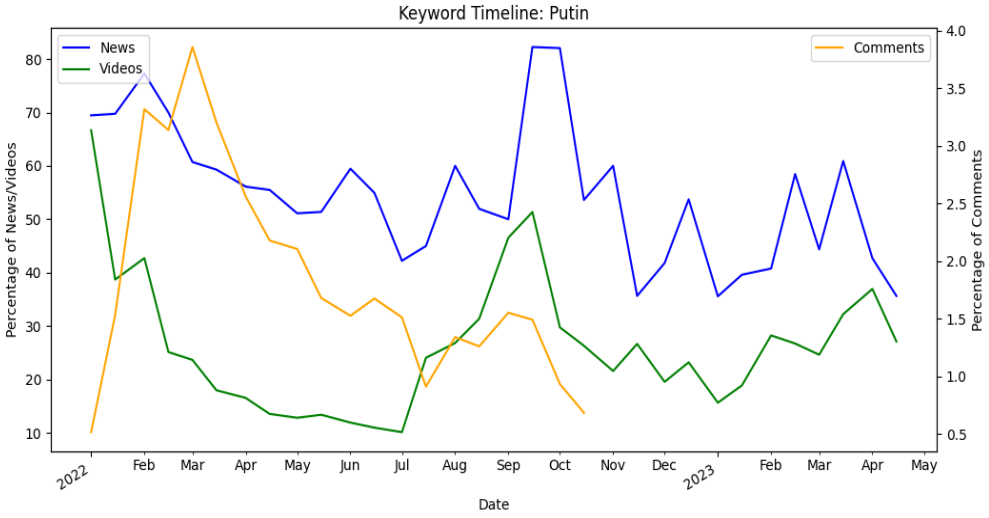


Figure 3: Timeline of mentions of "putin" in News vs TikTok videos

Furthermore, when the proportion of TikTok videos about a specific topic increases with the proportion of news articles, we find that many popular videos using the keyword do not actually discuss the specific topic. For example, use of the keyword "Odesa" increases in TikTok videos in tandem with news coverage, yet almost none of the videos we watched were related to the city of Odesa, but rather used the hashtag in large blocks of Ukraine-related hashtags, presumably to generate more visibility for their content. Although there may be many informational videos on TikTok about specific events that correlate with the timing of news coverage, these videos do not seem to receive as much engagement as those that take a clear and specific stance about the Russian-Ukraine war, with little additional explanation or connection to specific events.

We also calculate the Pearson correlation coefficient for the percentage of news and videos containing each keyword over time. Of eighteen keywords, nine have a coefficient greater than 0.3 and a p-value less than 0.05, demonstrating at least a moderate correlation between news and video content for half of the topics. Overall, we find some evidence that TikTok content about the major political event of the Russian-Ukrainian war is tethered to news coverage, suggesting that traditional news media may play a moderate, but very limited, role in influencing the dissemination of content on the platform.

3.2. Comparing comments and news

In contrast to TikTok videos, comments appear to engage more with specific news events. Of the eighteen keywords we focused on, fifteen show some simultaneity between comments and

news mentions. Additionally, we noticed an interesting phenomenon: in many cases where comments and news dynamics do not line up, comments about a specific topic will sharply increase shortly after the increase in news. To test this 'lag effect', we calculate Pearson correlation coefficients for the proportion of news and comments containing each keyword over time, allowing us to approximate the relationship between an increase in news and an increase in comments. We then calculate the correlation coefficient again, but this time offset the news proportions back by two weeks, effectively measuring the correlation between the proportion of news coverage two weeks prior and the proportion of comments.

In twelve out of eighteen topics, the lagged correlation coefficient is higher than the un-lagged correlation coefficient. For seven keywords, the coefficient is above 0.3 and the p-value is less than 0.05, demonstrating a notable moderate-to-strong correlation between the proportion of news two weeks prior and comments ("ukraine", "zelenskyy", "crimea", "nato", "denazification", "mariupol", "kherson"). In other words, how much attention traditional news media paid to a topic two weeks ago can sometimes be a better predictor of what people are currently discussing in the TikTok comment section than what is currently being covered in the news. Notably, we do not see a similar lag effect as often when comparing news and videos (seven topics with higher lagged coefficients, three with a significant p-value), suggesting that traditional news coverage may have a greater impact on conversation (i.e., comments) on TikTok than on videos themselves. Overall, we see frequent correlation between previous news coverage and comment discussion, suggesting that TikTok users are inspired by news coverage in their conversations about politics.

By reading a small sample of comments from instances in which topic mentions notably increase, we see that many TikTok users engage in real political discussion. For example, from March to June 2022, we see TikTok commenters discussing the labeling of Russia's invasion of Ukraine as a "denazification", although videos use the term "denazification" for a shorter period of time. Comments range from pointing out that Zelenskyy is Jewish ("They keep saying 'denazification' when the Ukrainian president is Jewish? Like what??") to re-iterating Putin's justification for the invasion ("Ukraine was killing its own citizens. Russia stepped in because no one else would. denazifying.").

Commenters also discuss previous conflicts relevant to the ongoing war, such as the 2014 conflict in Donbas ("ukrainians were bombing donbas for 8 years straight bro" "Russia bombed donbass not ukraine"), the annexation of Crimea ("We gonna talk about how Russia illegally occupies Crimea?", "Not true crimea was always Russian and filled with Russians bro in 2014 crimeans celebrated when Russia liberated them") and the 1999 Chechen War ("That's a well known fact even in Russia, that FSB did it by themselves, to start Second Chechen War. Read a book "Blowing Up Russia" by Y. Felshtinsky"). Commenters also mention specific events in the war to argue their larger political arguments. For example, Donetsk is often referred to in discussions about the region's Russian-backed independence movement ("what about the

freedom of luhansk Donetsk and Crimea who all voted to leave Ukraine but Ukraine won't acknowledge their independence"). Although the comments are not always backed up by fact, they do demonstrate that many TikTok users have knowledge of current and past news events and are well-informed (if not biased and/or misled) about the war.

Although our method allows us to identify how news coverage informs commenting behaviour on TikTok, as well as what topics are prominent among commenters, it is important to note that the proportion of comments that discuss a specific keyword in English is quite small. Many comments are in Ukrainian or Russian, and the most frequent English-language terms in the dataset of comments are quite generic, ranging from "ukraine" and "russia" to simply "ok", "oh", and "wow". So, although our analysis of individual TikTok videos and comments revealed more engagement with news events in the comment section than in videos themselves, these political discussions do not represent the majority of comments on political TikTok videos.

3.3. Comparing videos and comments

Finally, there are some instances where increases in TikTok video content about specific topics occur at the same time as comments about that topic, but the two are not always correlated. In other words, a TikTok video's description is not a good predictor of what is happening in the comment section. Based on the lack of videos referring to specific events/topics, this temporal correlation is likely due to the use of large amounts of hashtags in video descriptions to increase viewership. Content creators may be seeing that users discuss specific events in the comment section, inspiring them to add hashtags like #odesa or #azov to their videos, although the video itself is not actually about Odesa or the Azov Battalion.

4. Discussion

In our exploration of content about the Russian-Ukrainian War on TikTok, we observe two unique phenomena regarding the role of news on the platform. First, we find that, although TikTok and news coverage about specific topics sometimes line up, TikTok videos themselves do not frequently engage with news content. TikTok video creators – at least those that generate high levels of engagement – seem to be rarely inspired by specific news events to create their content, and instead primarily post "slice-of-life" or generally related content. TikTok creators appear to primarily use specific events as hashtags to gain visibility for their videos, while the content of the video serves to discuss a related or more general topic (e.g., using the hashtag "odesa" when the video is just broadly expressing support for Ukraine).

Second, TikTok users mainly engage with specific political events in the comment section, showing a slightly greater simultaneity with news coverage than videos and much stronger engagement with the topics themselves. We also observe a lag effect: for some topics, we observe a strong correlation between an increase in comments about the topic and an increase

in news coverage two weeks before, potentially demonstrating that many TikTok commenters actually do consume traditional news media and wait to get all the facts before engaging in conversation about a given topic. This may confirm findings by other researchers on the relationship between social media and news consumption. For example, Martin and Sharma (2022) find that people who get their news from social media also often get their news from real news outlets, so engagement with political content on digital platforms often serves to "accompany and complement traditional news use," (293). Many TikTok commenters are likely knowledgeable about specific news events, and use the comment section to express their views on them and engage in political discussion with others.

This study has implications for future work on the dynamics of political information on TikTok. Our primary contribution is the development of a methodology for studying engagement with news on TikTok. A platform like TikTok simply has too much content about any given topic, no matter how specific, and it is difficult for researchers to understand how information ebbs and flows on TikTok without knowing where to look. Our methodology enables researchers to analyze the relationship between news and content on TikTok without relying solely on correlation measures, allowing them to identify specific times where political conversation is particularly ripe on the platform and create a subset of videos or comments from that period without needing to depend upon TikTok's search interface, which prioritizes the most-viewed videos regardless of when they were posted. This comparison opens the door to further research about how TikTok users engage with news, how credibility and trust mediate the flow of information on the platform, and the evolving role that traditional news media has in online spaces.

References

Aslam, S. (2024). TikTok by the Numbers: Stats, Demographics & Fun Facts. Omnicore Agency.

Boeker, M., & Urman, A. (2022). An empirical investigation of personalization factors on TikTok. In Proceedings of the ACM web conference 2022 (pp. 2298-2309).

Literat, I., & Kligler-Vilenchik, N. (2021). How popular culture prompts youth collective political expression and cross-cutting political talk on social media: A cross-platform analysis. Social Media+ Society, 7(2), 20563051211008821.

Martin, J. D., & Sharma, K. (2022). Getting news from social media influencers and from digital legacy news outlets and print legacy news outlets in seven countries: The "more-and-more" phenomenon and the new opinion leadership. Newspaper Research Journal, 43(3), 276-299.

Steel, B., Parker, S., & Ruths, D. (2023). Invasion of Ukraine Discourse on TikTok Dataset. CoRR.

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17788

Exploring Enotourism's Impact on Winery Competitiveness through Online Data

Jose Baixauli[1] , Ana María Debon[2] , Roberto Cervelló-Royo[1], Josep Domenech[1]
[1]Department of Economics and Social Sciences, Universitat Politècnica de València, Spain [2]Centro de Gestión de la Calidad y del Cambio, Universitat Politècnica de València, Spain.

How to cite: Baixauli, J.; Debon, A. M..; Cervelló-Royo, R.; Domenech, J. 2024. Exploring Enotourism's Impact on Winery Competitiveness through Online Data. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17788

Abstract

Enotourism has become an increasingly popular form of tourism in the wine industry, offering visitors the opportunity to discover the wine culture, history, and production process while enjoying the natural and cultural heritage of the region. Enotourism can improve the wine brand image, heritage, cellar reputation, and the Denomination of Origin (DO) recognition. This paper explores how enoturism offers, including visits, tastings, restaurants, and accommodation services, relate to the competitiveness of wineries in Spain. To this end, financial variables obtained from the Sistema de Análisis de Balances Ibéricos (SABI) are combined with indicators extracted from the wineries' websites. PCA is applied to financial variables and the first three components are used to measure the different dimensions of winery competitiveness. Results show that each dimension is associated with a different set of concepts related to enoturism offer.

Keywords: *Digital footprint; web scraping; competitiveness; supervised learning; wineries; hospitality.*

1. Introduction

Nowadays, enotourism is a popular form of tourism in the wine industry, offering visitors the opportunity to discover the wine culture, history, and production process while enjoying the natural and cultural heritage of the region increasingly popular. In addition, enotourism can improve the wine brand image, heritage, and cellar reputation; well-known Spanish tourist destinations use wine as an essential attractor since the hospitality industry and wine are closely intertwined (Martínez-Falcó, 2023). Despite these findings, little literature exists on tourism, hospitality, and wine (Bonn, 2018). As Bianchi (2015) states, trust in a wine brand stems from wine experience and influences satisfaction with the brand. Wine tourism within the experiential economy encompasses a distinct lifestyle (Min, 2013), with the thematic focus of activities

being the most memorable factor in the overall experience (Saayman and Van Der Merwe, 2015). Moreover, the variety of cellars and the availability of complementary services within their surroundings play a significant role in determining the spectrum of activities to be developed (Bruwer and Alant, 2009). The types of services provided directly impact the wine experience (Carlsen, 2007). Thus, this paper explores how hospitality services, including visits, tastings, restaurants, and accommodation services on the websites, are related to the competitiveness of wineries at both individual and DO levels in Spain.

Our theoretical framework explores the multifaceted experiential model of enotourism. Thus, our model is multidimensional in nature whose fundamental pillars are the following; 1) Enotourism and visit itself (Wu and Liang, 2020; Priilaid, 2020), 2) Restauration and Hospiltality offer (Liu et al., 2017; Agyeiwaah et al., 2019), 3) Environment (Rachao et al., 2019; Madeira et al., 2019a, 2019b), 4) History and Heritage (Mason and O'Mahony, 2007; Fernandez and Cruz, 2016), 5) Protected and region designation of origin and/or protected geographically indication (Bianchi, 2015), 6) Cellars, vineyards, infrastructure and architecture (Cohen and Ben Nun, 2009), and 7) Sensorial Experience and wine tasting (Martins et al. 2017; Chen et al. 2016; Carlsen 2004, 2007).

The study uses a quantitative approach, combining data from different sources. A sample of 561 wineries located in Spain and their competitiveness variables are gathered from the SABI[1] database, which contains the balance sheets of companies operating in the wine industry. The website contents were automatically explored to identify the enoturism services and classify the dimensions of the offer by each cellar. Multivariate statistical and machine learning techniques (Hastie et al., 2009) are used to conduct a no-supervised analysis to discover patterns of competitiveness and subsequently supervised analysis to confirm the relation of these patterns with web content.

The results suggest that the increased level of enotourism positively influences the competitiveness of wineries. To extract the level of enotourism we chose and pulled the number of appearances of related keywords in the web. The competitiveness analysis reveals that companies that promote more enoturism services show better financial performance. Moreover, the content analysis of websites indicates that several wineries offer a range of services, including tours, tastings, restaurants, and accommodation.

[1] SABI stands for Sistema de Análisis de Balances Ibéricos (Iberian Balance Sheets Analysis System). It is a database published by Bureau van Dijk Electronic Publishing (BvDP), a Moody's Analytics company.

2. Material and Methods

2.1. Data

The study employs a data repository comprised of 4139 observations representing individual wineries in Spain. Each observation is uniquely identified by the winery's name and BvD number (SABI identifier). The database encompasses 16 variables categorized into two distinct groups: 10 economic variables and 6 digital footprint enoturism measures.

Economic variables were extracted from the SABI database, adhering to the specific variables employed by Castro et al. (2023). From 2019 to 2022 we extracted the: revenue, employees, assets, return on assets (ROA), return on equity (ROE), market share, value added, profit, value added per employee and revenue per employee. Only 1260 wineries were left after eliminating companies with missing values. With the remaining the mean was calculated.

To measure the dimensions of enotourism offer, we employed a methodology centered around analyzing data from company websites. Irrelevant content was filtered out, such as domains that had been hijacked or pages that were not possible to do web scrapping. Our focus remained on identifying websites specifically related to wines. From the 1260 wineries only 561 wineries offered some type of enoturism in their websites. Following this refinement, websites offering enotourism experiences were identified by employing a set of predetermined keywords derived from our theoretical model. These keywords were stemmed and used to search through the website contents, and the keywords were quantified within different categories of enotourism. This allowed for the calculation of the percentage representation of these keywords across the enotourism categories.

2.2. Multivariate analysis

Statistical analyses were conducted using the R environment for statistical computing (R Core Team, 2023). Principal component analysis (PCA) was used to identify outliers (Ferrer-Lorenzo et al. 2018) and to obtain measures of business competitiveness (Dess and Davis, 1984; Ibrahim et al., 2001; Ortega, 2010; Ferrer-Lorenzo et al. 2018). Subsequently, digital footprint indicators were used to create a regression model to explain the different components of the winery competitiveness.

3. Results

After removing outliers, PCA was applied to the competitive variables extracted from SABI. From the PCA results the first 3 components were selected as they all had an eigenvalue greater than 1 and explained a total of nearly 80% of the variance (Figure1).

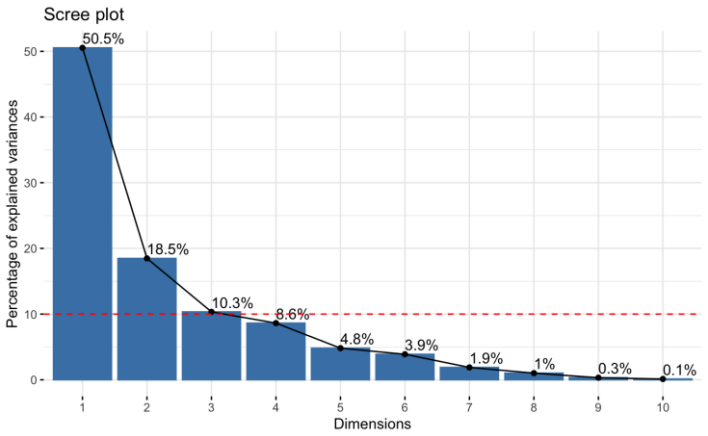


Figure 1. Percentage of explained variance of each component

After analyzing the contributions of variables to of the first two components (Figure 2), we designated the first component as the "Size" because it was positively correlated with variables such as revenue (Rev), number of employees (Empl), value added (VA), market share (MS), and total assets (Assets). In contrast, the second component predominantly featured relative variables like return on assets (ROA), value added per employee (VA_Empl), and revenue per employee (Rev_Empl), leading us to identify it as reflective of the company's "Productivity." Lastly, the third component was notably linked to return on equity (ROE), prompting us to label it as "Profitability."

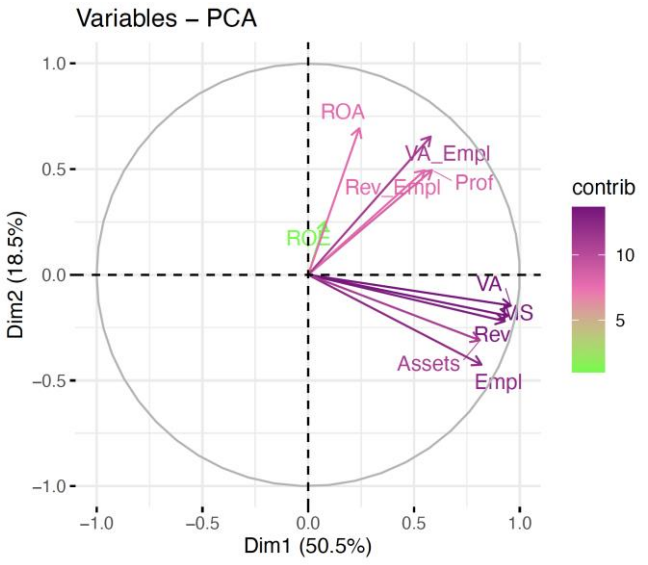


Figure 2. Variables contribution to first and second component

We employed a regression linear model to discover the impact of various types of enotourism contents on company websites on their competitiveness (Table1). Our analysis revealed that marketing initiatives focused on visits and the historical aspects of enotourism are positively associated with company size. These may occur because marketing initiatives that highlight visits and historical aspects of enotourism can attract a larger audience interested in wine and cultural heritage. Therefore, larger wineries tend to emphasize different enoturism activities and the visit itself.

Infrastructure exhibited a negative association with productivity. When a winery invests heavily in infrastructure, it often means they provide a wide array of services such as boutiques, terrace gardens, and more. While these services may enhance the customer experience, they could also divert resources and attention away from the primary goal of wine production. As a result, the winery may become less efficient in its core operations. This inefficiency can manifest as lower productivity, hence the negative association of excessive infrastructure.

Similarly, the emphasis of infrastructure is negatively associated with profitability. However, we also saw that enhancing the sensorial experience for visitors yielded a significant positive effect on overall profitability. This is likely because a superior sensorial experience can create a more memorable and engaging customer visit. Customers with a unique and enjoyable experience are more likely to purchase products, return for future visits, and recommend the winery to others. This positive association can increase sales, customer retention, and word-of-mouth marketing, contributing to higher profitability.

4. Conclusions

Enotourism is an established trend that significantly impacts wineries. This study has examined the characteristics associated with enotourism presented on their websites and its correlation with the competitiveness of the wineries.

Various aspects of competitiveness are intrinsically linked with distinct characteristics of enotourism. A clear association was observed between the enoturism promoted and the competitivity of the wineries. Larger wineries tend to emphasize the promotion of historical facets and touristic visits. Conversely, wineries with lower productivity often exhibit a diverse range of infrastructure. Most notably, the most profitable wineries strategically focus on delivering a sensory experience to their customers.

References

Agyeiwaah, E.; Otoo, F.E.; Suntikul, W.; Huang, W.J. Understanding culinary tourist motivation, experience, satisfaction, and loyalty using a structural approach. J. Travel Tour. Mark. 2019, 36, 295–313.

Table 1. Effect of Enoturism on Competitiveness. Standard errors in parentheses. Dependent variables in logarithmic form. *** p < 0.001; ** p < 0.01; * p < 0.05

	Size	**Productivity**	**Profitability**
visit	0.461***	0.054	0.054
	(0.137)	(0.071)	(0.071)
hospitality	-0.088	-0.031	-0.032
	(0.162)	(0.084)	(0.084)
history	0.256*	0.030	0.029
	(0.123)	(0.064)	(0.064)
DO	0.208	-0.052	-0.052
	(0.147)	(0.076)	(0.076)
infrastructure	0.176	-0.244*	–0.157**
	(0.192)	(0.100)	(0.050)
sensorial exper	-0.232	0.016	0.107**
	(0.126)	(0.066)	(0.033)
(Constant)	0.930 ***	2.016***	2.690 ***
	(0.068)	(0.035)	(0.018)
N	561	561	561
R^2	0.124	0.033	0.034

Bianchi, C. (2015). Consumer brand loyalty in the Chilean wine industry. Journal of food products marketing, 21(4), 442-460.

Blazquez, D., Domenech, J., Debón, A. (2018). Do corporate websites' changes reflect firms' survival? Online Information Review, 42(6), 956-970.

Blazquez, D., Domenech, J. (2018). Web data mining for monitoring business export orientation. Technological and Economic Development of Economy, 24(2), 406-428.

Bonn, M. A., Cho, M., Um, H. (2018). The evolution of wine research: A 26 year historical examination of topics, trends and future research direction. International Journal of Contemporary Hospitality Management, 30(1), 286-312.

Bruwer, J.; Alant, K. (2009) The hedonic nature of wine tourism consumption: An experiential view. Internationa Journal of Wine Business Research , 21, 235–257.

Castro, L.; Debón, A.; Domenech, J (2023). Study of the relationship between competitiveness and digital footprint indicators in Valencian wineries. In: 5th International Conference on Advanced Research Methods and Analytics (CARMA 2023). Sevilla, 28-30 June 2024. https://doi.org/10.4995/CARMA2023.2023.17009

Carlsen, P.J. A review of global wine tourism research. J. Wine Res. 2004, 15, 5–13.

Carlsen, J. (2007) Global Wine Tourism: Research, Management and Marketing; CABI:Wallingford, UK.

Cohen, E.; Ben-Nun, L. The important dimensions of wine tourism experience from potential visitors' perception. Tour. Hosp. Res. 2009, 9, 20–31.

Dess, G.G., Davis, P.S., (1984). Generic strategies as determinants of strategic group membership and organizational performance. Academy of Management Journal 27 (3), 467–488.

Fernandes, T.; Cruz, M. Dimensions and outcomes of experience quality in tourism: The case of Port wine cellars. J. Retail. Consum. Serv. 2016, 31, 371–379

Ferrer-Lorenzo, J. R., Maza-Rubio, M. T., Abella-Garcés, S. (2018). The competitive advantage in business, capabilities and strategy. What general performance factors are found in the Spanish wine industry?. Wine Economics and Policy, 7(2), 94-108.

Hastie, T., Tibshirani, R., Friedman, J. H., (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.

Ibrahim, B., Dumas, C., McGuire, J., (2001). Strategic decision making in small family firms: an empirical investigation. Journal of Small Business Strategy 12 (1), 80–90.

Liu, W.; Sparks, B.; Coghlan, A. Event experiences through the lens of attendees. Event Manag. 2017, 21, 463–479.

Martínez-Falcó, J., Marco-Lajara, B., Zaragoza-Sáez, P. y Sánchez-García, E. (2023) Wine tourism in Spain: The economic impact derived from visits to wineries and museums on wine routes. Investigaciones Turísticas (25), pp. 168-195. https://doi.org/10.14198/INTURI.21219

Martins, J., Gonçalves, R., Branco, F., Barbosa, L., Melo, M., & Bessa, M. (2017). A multisensory virtual experience model for thematic tourism: A Port wine tourism application proposal. Journal of destination marketing & management, 6(2), 103-109.

Madeira, A.; Correia, A.; Filipe, J.A. Modelling wine tourism experiences. Anatolia 2019a, 30, 513–529.

Madeira, A.; Correia, A.; Filipe, J.A. Wine Tourism: Constructs of the Experience. In Trends in Tourist Behavior; Springer: Berlin/Heidelberg, Germany, 2019b; pp. 93–108.

Mason,R.; O'Mahony, B. On the trail of food and wine: The tourist search for meaningful experience. Ann. Leis. Res. 2007, 10, 498–517.

Min,W (2013). Analysis of the wine experience tourism based on experience economy: A case for Changyu wine tourism in China. Res. Journal of Applied Science Engineering.Technology 2013, 5, 925–4930.

Ortega, M.J., (2010). Competitive strategies and firm performance: technological capabilities' moderating roles. Journal of Business Research 63 (12), 1273–1281.

Priilaid, D.; Ballantyne, R.; Packer, J. A blue ocean strategy for developing visitor wine experiences: Unlocking value in the Cape region tourism market. J. Hosp. Tour. Manag. 2020, 43, 91–99.

Rachao, S.; Breda, Z.; Fernandes, C.; Joukes, V. Food and wine experiences towards co-creation in tourism. Tour. Rev. 2019, 76, 1050–1066.

Saayman, M.; Van Der Merwe, A. (2015) Factors determining visitors' memorable wine-tasting experience at wineries. Anatolia, 26,372–383.

Wu, G.; Liang, L. Examining the effect of potential tourits' wine product involvement on wine tourism destination image and travel intention. Current Issues Tourism 2020, 1–16.

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17795

In What is Europe Investing? A Text Mining Approach on Cohesion Projects

Nicola Caravaggio, Giuseppe Di Renzo, Laura Fanelli, Giuliano Resce, Agapito Emanuele Santangelo
Department of Economic, University of Molise, Italy.

How to cite: Caravaggio, N.; Di Renzo, G.; Fanelli, L.; Santangelo, A. E. 2024. In What is Europe Investing? A Text Mining Approach on Cohesion Projects. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17795

Abstract

In this Paper we analyze the dynamics of the interventions of European cohesion policies in the Italian territory - according to a regional division - focusing on the topics financed, extracted through text mining, in particular with the term document matrix and term frequency – inverse document frequency. From the analysis conducted on 51.971 projects financed from 2000 to 2022, a clear preponderance of the topics of internships, civil service and more generally aspects relating to the world of work emerges. Although these are the most frequent topics, by weighing the frequency of words in the summary of the call with the total value of the financing, it emerges that the most costly projects in terms of resources are those linked to the development of infrastructure. This innovative approach allows an innovative understanding of the spending trends, useful to improve the action of the European political decision- maker.

Keywords: *Cohesion Policy; Text mining; Public investments; Term Document Matrix; Term frequency – inverse document frequency*

1. Introduction

The European Cohesion Policy is a series of interventions and measures aimed at reducing economic and social inequalities among territories, thus promoting homogeneous, inclusive, and long-term growth. The Cohesion Policy is provided for by art. 119 of the Italian Constitution and the Treaty on the Functioning of the European Union. Cohesion Policy essentially finds its genesis in the very definition of the European Union, or rather from the observation that the creation of a single market, as well as a regulation valid for all member states, would inevitably have aggravated the disparities and differences in terms of per capita income and wealth of the territories. Cohesion interventions involve central authorities and local administrations equally. When evaluating the impact of Cohesion Policies, it is necessary to focus more on the outputs

of the interventions, in terms of reducing economic-social inequalities, rather than on the quantity of resources used, which represent only the amount of expense.

But, in the study of the long-term growth process it is believed that capital accumulation is not enough to guarantee continuous and sustainable growth in the long term. Neoclassical economic analysis believes that, alongside the accumulation of capital, a driver of growth is agglomeration, and therefore the movement of large masses of individuals from rural and/or peripheral areas towards metropolitan centers, favoring thus the contamination of knowledge, ideas, and therefore enrichment of human capital and growth of technological progress (Marshall, 1890) The empirical evidence found in recent years highlights a dichotomous picture: dynamic urban agglomerations and remote regions characterized by low growth rates and depopulation (Iammarino et al., 2019). From this, it follows, as a logical corollary, that the colorful European picture is composed on the one hand of prosperous nations (central-northern Europe), while other areas are characterized by low growth and/or low income (southern Europe, Eastern Europe). Considered analytically, even the leading States in terms of wealth and economic growth present internal conditions of imbalance also in terms of wealth distribution.

Therefore, in the analysis of growth processes it is impossible to ignore the consideration that in Europe many regions have very different growth paths, compared to the performances of the States to which they belong (Cuaresma et al., 2014). This means that a prosperity of a Nation does not necessarily determine homogeneous prosperity across its entire territory (Liberati et al., 2022). Rather, it is the local, atomistic dynamics that guide the progress (Petrakos et al., 2011), and therefore the analysis of regional gaps and internal dynamics constitutes a crucial point for adapting cohesion interventions to the emerging needs of the territories, which are increasingly attentive to attributes such as sustainability, respect for the environment, the protection of workers' rights.

However, if today we still continue to debate the effectiveness of the EU's cohesion policy – despite the fact that its importance has become clear recently, during the pandemic period (European Commission, 2022) – it means that, not considering merely ideological contestations, there are problems that do not allow cohesion interventions fully express their potential. On the one hand, its detractors point out that, despite the enormous amount of resources used in recently years, the disparities have worsened only modestly, and that in reality, the Cohesion Policy only serves to "keep public employees and resources busy", but at the same time preventing them from carrying out and carrying out more productive activities (Molle, 2007); others, however, believe that the cohesion interventions have actually achieved good results (Gagliardi & Percoco, 2016). The truth is probably somewhere "in the middle": Cohesion policy is important, but it still suffers from a certain delay in its implementation due to the combined effect of various factors, such as the inability to plan and spend these resources. Regardless of these considerations, one fact remains constant, and remains a topos in this context, at least as regards

the levels of effectiveness of the Cohesion Policy and the investments related to it: the low quality of the institutional context, a determining factor in guaranteeing the operation of the convergence and growth mechanisms of the European Regions. In particular, the ability of local political decision-makers proves crucial (Fratesi & Whishlade, 2017) in ensuring that large investments produce the corresponding fruits (Arbolino & Boffardi, 2017). The latter, in particular, are crucial to allow the so-called less developed Regions to become Regions in transition and subsequently, hopefully, more developed. According to the data provided by OpenCoesione, for the 2014-2020 programming cycle, taking into account the Italian framework, only 13% of the projects were concluded. The data is striking, and prompts reflections especially regarding the implementation time of public works in Italy. In a specific report from the Agency for Territorial Cohesion, the figure of the slowness regarding the construction and completion of public works emerges.

The average implementation time of the infrastructure works is approximately 4.4 years, but progressively increases as the economic value of the projects increases. These timings are necessarily also influenced by the reference contexts.

This problem emerges, naturally, for those European areas, such as southern Italy, where in the last twenty years low growth rates have been recorded, if not close to zero or negative as in the case of real wages, also deriving from the fact that investments carried out have been victims of the so-called "distributive drift", with a consequent excessive fragmentation of the interventions and therefore also of their unitary dimension (Agrello, 2019). This favors a sort of sub-optimization phenomenon: the management of resources by local political decision-makers - from a realistic and rational point of view interested in their reconfirmation and therefore in maintaining their office - will inevitably be directed towards political goals, favoring only the preservation of the status quo, without any real prospects for change and modernization being revealed. From this perspective - while distancing ourselves from any form of demonization of the work of local politicians

- in order to bring about a real positive turning point, it would be appropriate to disconnect the operation of cohesion policies from the political cycle and from the purely electoral logic, through actions long-term (such as improvements in education) and the provision of tools to support the action of local public administrations.

Despite the increase in monitoring initiatives for cohesion policies, the majority of indicators primarily focus on expenditure, with limited evidence attempting to understand the actual impact of the projects. This paper aims to address this gap by employing text mining techniques to analyze all European projects financed in Italy from 2000 to 2022 Preliminary results offer fresh insights into the priorities targeted by cohesion funding, paving the way for a novel approach to policy analysis.

2. Methods

Advances in machine learning research offer great potential for international development agencies to leverage the vast information generated from accountability mechanisms to gain new insights, providing analytics that can improve decision-making. (Resce, Garbero & Carneiro, 2021)Within digitally based content analysis approaches, text Mining is broadly defined as an Artificial Intelligence (AI) technique that uses Natural Language Processing (NLP) to transform unstructured text of documents/databases such as web pages, newspaper articles, e-mails, fundings, press, posts/comments on social media, in structured and normalized data (Resce & Maynard 2018). Words, the carriers of meaning, are identified and transformed into a processable data structure.

Indeed, many studies focus on this technique to provide an aid to policy maker to optimally allocate resources: Resce et al, (2021) explores how text mining can harness existing project data to uncover latent information about food systems dynamics taking 900 projects of the International Fund for Agricultural Development (IFAD); Choudhary et al (2009), instead, identified text mining as a potential tool for addressing the identified problems of Post-Project Reviews.

In our study, text mining has been employed to analyze 50.971 projects implementing cohesion policies from 2000 to 2022, funded by Structural Funds, the National Fund for Development and Cohesion (FSC), and the Cohesion Action Plan (PAC). Our objective is to identify predominant themes in the projects and analyze their evolution over the years, with particular attention to the division into the macro-areas of Italy, namely "Centre-North" and "South". (Open Coesione, 2023)

The first step in the analysis involved retrieving data, including the project code, title, summary, destination macro-area, funding amount, and project start date. The text chosen for analysis was the summary of each project, which was extracted and aggregated based on the project's start year and month. Subsequently, the analysis corpus was prepared using functions from the R package "tm" (Feinerer & Hornik, 2018; Feinerer, Hornik & Meyer, 2008): punctuation, stop words (e.g. words like "the," "is," "of," etc.), and numbers were removed from the corpus. The words were then converted to lowercase and stemmed. Finally, a Term Document Matrix was produced, with the project's start year and month as columns (126) and words as rows (28 928 unique terms). The Term Document Matrix indicates the number of times each word appears in each project started in that month and year. It serves as the starting point for text mining by transforming unstructured text into numerical data.

A central question in text mining is how to quantify what a document is about. One measure of a word's importance is its term frequency (tf), which counts the occurrence of a word in a document. Another approach is to consider the inverse document frequency (idf), which decreases the weight of commonly used words and increases the weight of words that are not

frequently used in a collection of documents. The two can be combined to calculate the tf-idf of a word (the product of the two quantities), which measures the frequency of a word adjusted for how rarely it is used (Silge & Robinson, 2017). Formally:

$$\text{idf(term)} = \ln\left(\frac{n_{\text{documents}}}{n_{\text{documents containing term}}}\right) \quad (1)$$

The statistic tf-idf is widely used to measure how important a word is to a document in a collection of documents (Silge & Robinson, 2017). In our case, the tf-idf combines frequency - how many times a word is associated is associated to month and year of the project's start - and the inverse of ubiquity - how exclusive the association is between a word and month and year of the project's start. To this regard, it is worth stressing that more ubiquitous words are more likely to have less informative power than exclusive words.

The Term Document Matrix and tf-idf statistics were initially computed for all projects and later differentiated for projects targeted at the "Center-North", composed by twelve regions (Liguria, Lombardia, Piemonte, Valle d'Aosta, Emilia-Romagna, Friuli Venezia Giulia, Trentino-Alto Adige, Veneto, Lazio, Marche, Toscana and Umbria) and "South" of Italy, so called "Mezzogiorno" composed by eight regions (Abruzzo, Basilicata, Calabria, Campania, Molise, Puglia, Sardegna and Sicilia). Using the functions from the R package "wordcloud2", project keywords were incorporated into word clouds, visual representations where a word's size is proportional to its frequency (Silge & Robinson, 2017).

Subsequently, graphs were generated to illustrate the trend of words over time. Finally, to comprehend the economic value associated with the most frequent words in the projects, a new matrix was created by weighting the Term Document Matrix according to the project's funding value where the word is present.

3. Results

Among the 50.971 projects implemented under cohesion policies from 2000 to 2022, the findings in Figure 1 show that the most frequent words, using the Term Document Matrix and the Term Frequency-Inverse Document Frequency, are nearly identical

Figure 1. Comparison of the most frequent words in projects using the Term Document Matrix (wordcloud on the left) and the Term Frequency-Inverse Document Frequency (wordcloud on the right).

Most projects are dedicated to internships and civil service. Among the internships, non-curricular and extracurricular internships stand out, aimed at training young individuals for the world of work, followed by curricular internships targeting students to integrate professional training with academic education. Similarly, projects dedicated to civil service represent a significant opportunity for training and professional maturity. This highlights how cohesion policies prioritize youth growth and labour.

Examining Figure 2, the trend over time of the frequency of the top 10 words in the calls for proposals is evident, using both Term Frequency and Term Frequency-Inverse Document Frequency.

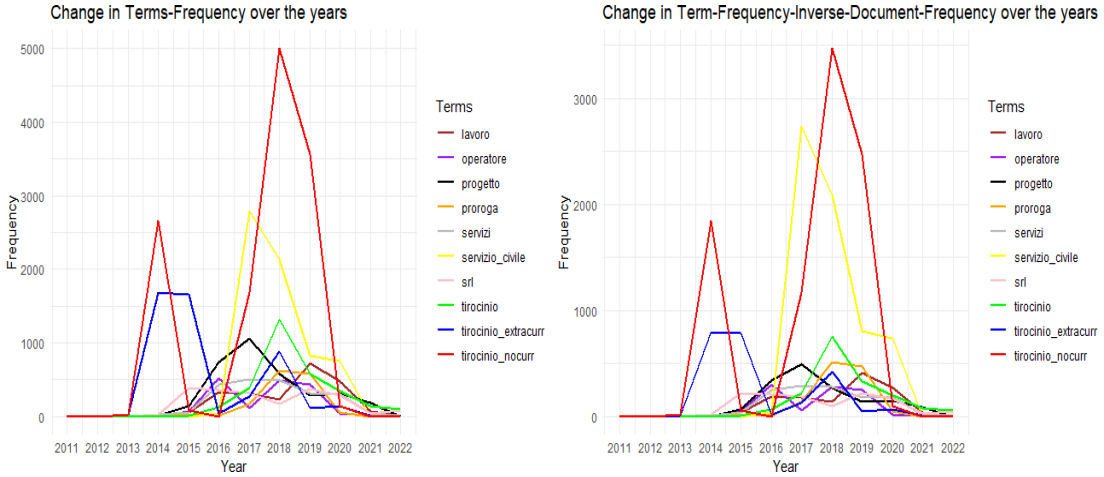


Figure 2. Variation in the frequency of the top 10 words over time from 2011 to 2022 using the Term Document Matrix (left plot) and the Term Frequency-Inverse Document Frequency (right plot).

The x-axis represents the project initiation year, while the y-axis represents the frequency. Two peaks are evident, the first in 2014 and the second in 2018, corresponding to the years when the majority of calls containing the most frequent words were implemented. It is noteworthy that from 2019 onwards, the lines began to decline, showing a trend towards flattening in the years 2020-2022, during the Covid-19 pandemic.

In this context, it is interesting to examine Figure 3, which focuses on the years 2020-2022 and illustrates how, although calls related to civil services continue to predominate, attention has shifted towards projects related to employment and occupation, as well as others addressing contributions and incentives.

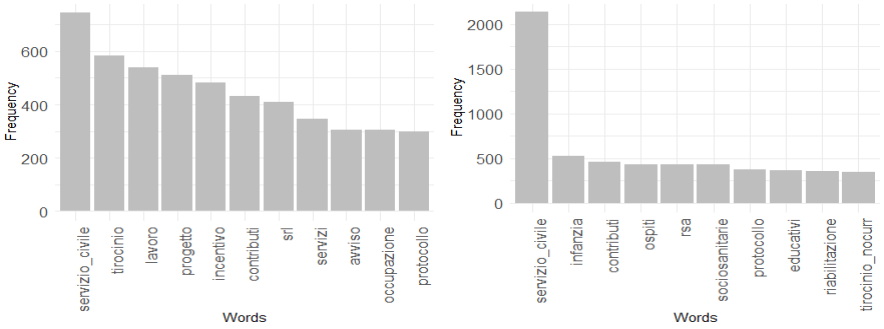


Figure 3. Frequency of the top 10 words from 2020 to 2022 using the Term Document Matrix (left plot) and the Term Frequency-Inverse Document Frequency (right plot).

As mentioned earlier, there exists an economic and social gap among Italian regions, confirmed by the number of initiated projects. In the North-Central area, there were 44 255 projects (86.82%), while in the South, there were only 6 716 projects (13.18%).

Figures 4 and 5 present the results of the repeated analysis, dividing the calls into the macro areas 'North-Central' and 'South'.

Figure 4. Comparison of the most frequent words in projects with the Term Document Matrix between the "Central_North" macroarea (left wordcloud) and the "Southern" macroarea (right wordcloud).

Figure 5. Comparison of the most frequent words in projects with the Term Frequency-Inverse Document Frequency between the "Central_North" macroarea (left wordcloud) and the "Southern" macroarea (right wordcloud).

While the most common words in the North-Central region coherently reflect what is illustrated in Figure 1, new key terms emerge in the South, such as “relocation”. “contract”, “well-being” and “welfare”. These highlight the ongoing challenge related to occupational crisis situations in the region. Indeed, one of the objectives of the calls is to mitigate the effects of business difficulties in the involved areas and promote the preservation of employment levels.

In Figure 6, the trend from 2011 to 2022 of the frequency of the most common words in the calls has been compared, dividing them by macro area.

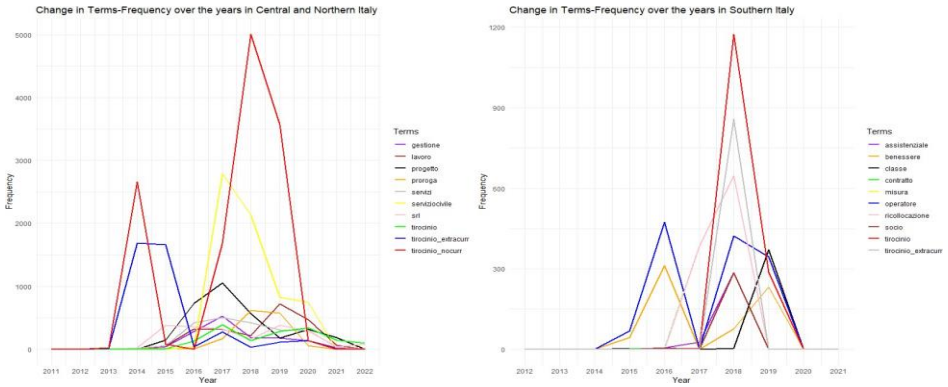


Figure 6. Comparison of change in frequency of the first 10 words over time from 2011 to 2022 with the Term Document Matrix between the "Central_North" macroarea (left plot) and the "Southern" macroarea (right plot).

In the North, the projects containing the most frequent words began in 2013, reaching a peak in 2014, followed by a period of contraction in 2016. In the South, they started later, in 2014, with a peak in 2016 and a sharp decline in 2017. Both macro areas experienced another peak in 2018. However, while in the North-Central region, a certain level of project initiation has been maintained even after 2020, even if in a smaller size; in the South, there has been a total

interruption of new projects. What happens from 2020 to 2022? While the left graph in Figure 7 confirms the previous observations about the Central- Northern region, the right graph displays the most frequent words in projects for Southern Italy. It becomes evident that the funding is primarily directed towards contributions to individuals for acquiring services, including training, with a particular focus on specific roles such as "esthetic treatment technician." The specific objective is to reduce the rate of early educational failure and scholastic and training dispersion.

The remaining words are related to projects launched for specific municipalities in the Sicily region, including "Catania," "Acireale," "Agrigento," and "Caltagirone," which are about social inclusion and the fight against poverty. These initiatives aim to increase, consolidate, and qualify care services and infrastructure.

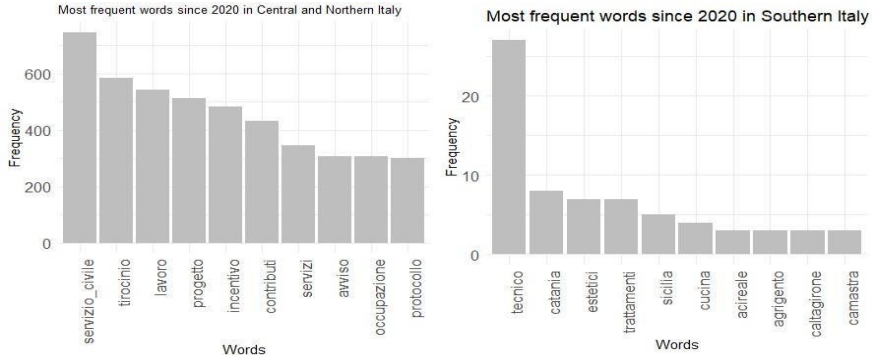


Figure 7. Frequency of the first 10 words from 2020 to 2022 with the Term Document Matrix between the "Central North" macro area (left plot) and the "Southern" macro area (right plot).

In returning to the comprehensive analysis of the calls, it is valuable to identify the words that exhibit a correlation with higher funding amounts. Let's focus on Figure 8, which considers not only the frequency of words but also the associated funding values.

Figure 8. Words in the tenders weighted by the value of European funding

Considering that cohesion policies aim to reduce territorial disparities, it is not surprising that among the projects with a more significant economic impact are those focused on the development of infrastructure, such as roads, ports, and railway networks. This helps us

understand that, although many projects focus on education and training (civil service, internships, etc.), the most substantial funding and truly impactful projects are directed towards the transport and mobility sector.

4. Conclusions

This study investigates what the projects implementing cohesion policies describe, through text mining analysis on administrative documents produced by local authorities.

Results shows how a huge part of funded projects focus on civil service and internships, hence they raise questions concerning economic policy and long-term development. While such initiatives can be effective in providing temporary support during periods of declining employment, it is critical to recognize that they primarily address the immediate consequences of unemployment rather than facing the structural causes that contribute to the declining employment itself.

Furthermore, the inappropriate use of civil service projects and internships may reflect a lack of strategic investment in innovation, education and professional training, which are crucial pillars for stimulating sustainable and inclusive economic growth the long term. Without adequate attention to these sectors there is a risk that national economic growth will remain limited, and society will continue to rely mostly on temporary measures to address employment challenges.

Additionally, it is important to consider the role of employment quality in the economy. While internships can offer work experience or a first entry into the job market for youths, if they are not accompanied by opportunities for professional and salary growth, they may not be able to provide a solid foundation for long-term work gratification and stability and for overall economic prosperity.

Therefore, although it is crucial to provide immediate support through civil service and internships, it is equally important to take a long-term perspective and invest resources and efforts in creating quality job opportunities, innovation, and professional training. Only through a balanced approach that addresses both immediate and long-term needs will it be possible to promote sustainable and inclusive economic growth.

References

Agenzia per la Coesione Territoriale (2018). Temi Cpt; Rapporto sui tempi di attuazione delle Opere Pubbliche. Dipartimento per lo Sviluppo e la Coesione Economica (2014). I tempi di attuazione e di spesa delle Opere Pubbliche.

Agrello, Pietro (2019). La politica di coesione: l'esperienza italiana. Rivista italiana di public management; Vol 2, n.1 (147-166). Marshall, A. (1890). "Principles of economics". Macmillan, London.

Arbolino, Roberta & Boffardi, Raffaele (2017). The Impact of Institutional Quality and Efficient Cohesion Investments on Economic Growth Evidence from Italian Regions. Sustainability 9, no. 8: 1432. https://doi.org/10.3390/su9081432.

Choudhary, A. K., Oluikpe, P. I., Harding, J. A., & Carrillo, P. M. (2009). The needs and benefits of Text Mining applications on Post-Project Reviews. Computers in Industry, 60(9), 728-740.

European Commission (2022). The 8th Cohesion Report. Molle, Willem (2007). "European Cohesion Policy".

Feinerer, I., & Hornik, K. (2018). tm: Text Mining Package. R package version 0.7-6. URL: https://CRAN. R-project. org/package= tm. Feinerer, I., Hornik, K., & Meyer, D. (2008). Infrastruttura di text mining in R. Journal of statistical software, 25, 1-54.

Fratesi, Ugo & Wishlade, Fiona G. (2017). The impact of European Cohesion Policy in different contexts. Regional Studies, 51:6, 817-821, DOI: 10.1080/00343404.2017.1326673.

Gagliardi, Luisa & Percoco, Marco (2016). The impact of European Cohesion Policy in urban and rural regions. Regional Studies; DOI: 10.1080/00343404.2016.1179384.

Garbero, A., Carneiro, B., & Resce, G. (2021). Harnessing the power of machine learning analytics to understand food systems dynamics across development projects. Technological Forecasting and Social Change, 172, 121012.

Iammarino, S., Rodrıguez-Pose, A., and Storper, M. (2019). Regional inequality in Europe: evidence, theory and policy implications. Journal of economic geography, 19(2):273–298.

Jesús Crespo Cuaresma, Gernot Doppelhofer & Martin Feldkircher (2014). The Determinants of Economic Growth in European Regions, Regional Studies, 48:1, 44-67, DOI: 10.1080/00343404.2012.678824.

Liberati, P., Resce, G., & Tosi, F. (2022). The probability of multidimensional poverty: A new approach and an empirical application to EU-SILC data. Review of Income and Wealth.

OpenCoesione. (2024). Projects with Extended Path. Retrieved February 9, 2023 from https://opencoesione.gov.it/it/opendata/#!progetti_section

Petrakos, G., Kallioras, D., & Anagnostou, A. (2011). Regional convergence and growth in Europe: understanding patterns and determinants. European Urban and Regional Studies, 18(4), 375-391. DOI: 10.1177/0969776411407809.

Resce, G., & Maynard, D. (2018). What matters most to people around the world? Retrieving Better Life Index priorities on Twitter. Technological Forecasting and Social Change, 137, 61-75.

Silge, J., & Robinson, D. (2017). Text mining with R: A tidy approach. " O'Reilly Media, Inc.".

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17801

Can websites reveal the extent and degree to which a business's values reflect national policy? A text embeddings approach

Alexander Hogan[1], Stephanie Cussans Moran[2], Kevin Hogan[3], Beth Barker[4], Richard Woodall[5]

[1]AI Research and Design, Etic Lab LLP, Wales, [2]Etic Lab LLP, Wales, [3]Etic Lab LLP, Wales, [4]Etic Lab LLP, Wales, [5]Etic Lab LLP, Wales.

How to cite: Hogan, A.; Moran S.C.; Hogan K., Barker B., Woodall R. 2024. Can websites reveal the extent and degree to which a business's values reflect national policy? A text embeddings approach. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17801

Abstract

This paper demonstrates a novel application of text embeddings and vectorisation to website texts to measure the prevalence and proliferation of values associated with the Well-being of Future Generations (Wales) Act 2015 amongst Welsh businesses. This was achieved by training a model to recognise topics of particular importance in texts describing the WFGA, and then using this model to analyse the text hosted on Welsh business's websites. The companies studied comprised a structured national sample of the Welsh economy. Our findings indicate that the method detects meaningful differences between regions, business sector and participants in a government innovation support program.

Keywords: *Econometrics; sociometrics; policy evaluation; big data analytics; text vectorisation; text embeddings.*

1. Introduction

Recent years have posed a series of pressing challenges for measuring business performance. Heightened social, political and economic volatility (including the pandemic, inflation and energy price shocks) have subjected economies and businesses to an accelerated rate of change, meaning that traditional survey-based business intelligence has lagged behind real-time rapid change. At the same time, an increased digitalisation of businesses driven by the pandemic has seen companies 'rapid adoption of remote working practices and increased online engagement. The increased frequency of digital activity creates an opportunity to develop robust digital business intelligence metrics that can be collected at scale in near real-time on an ongoing basis.

This paper presents a new approach to measuring the prevalence and proliferation of a particular set of values amongst a specific community of businesses. It describes how we trained a model

to measure the text on business websites against a unique piece of national legislation, the Well-being of Future Generations (Wales) Act 2015, using a sentence vectorisation method. We applied this to text scraped from a structured sample of company websites across Wales, representative by business type and unitary authority, in order to measure the extent and degree to which they represent the values put forward in the Act.

The question we aim to answer with this case study is whether our metric is capable of meaningfully discriminating across a sample. We measured this by looking at variations across unitary authority or by business sector (represented by SIC code), and variations registering the impact of a Welsh government business support program, within the sample. When tested against a structured sample of 5,431 Welsh companies, the results showed meaningful differences in the prevalence and contextual salience of values associated with the WFGA across regions, business type and participants in a government innovation support program.

Wales and the Well-being of Future Generations (Wales) Act 2015, herein the WFGA, represent an interesting test case for this technology. Firstly, Wales and the WFGA offer a well-defined, discrete entity; the Act perfectly overlaps with geographic area and the businesses we are measuring, as it applies to businesses across the whole of Wales. Moreover, the compact size of Wales makes it practical to build a representative sample of businesses by Unitary Authority (UA) that reflects the distribution of businesses in each UA by business sector (SIC code). Secondly, the WFGA requires government departments, businesses and other organisations to factor the health, environmental, sustainable, equitable and economic prospects of future generations into their decision making. That is to say, the WFGA requires a public response from government and businesses, that demonstrates they subscribe to the values and are trying to act on them. This can be measured more easily than, for example, carbon emissions or volume of waste recycled. Finally, it offers a case study of a regional and rural economy that differs from highly digitised and capitalised urban cosmopolitan regions, hence providing an interesting test for this methodology and whether it can produce meaningful results across urban and rural geographies with differing levels of digitisation (Wales Digital Maturity Survey 2023).

2. The WFGA

The WFGA is a unique piece of Welsh legislation, passed in 2015, which creates a legal obligation for:

"public bodies in Wales to think about the long-term impact of their decisions, to work better with people, communities and each other, and to prevent persistent problems such as poverty, health inequalities and climate change" (Future Generations Commissioner for Wales).

The aims of the act, expressed through seven "wellbeing goals", are: a prosperous Wales, a resilient Wales, a more equal Wales, a healthier Wales, a Wales of cohesive communities, Wales

of vibrant culture and thriving Welsh language, and a globally responsible Wales. There is an increasing prevalence of schemes that utilise government training and procurement processes to incentivise businesses to sign up to social projects such as net zero, community wealth-building and social value; the WFGA represents a unified, strategic attempt to achieve this on a national scale with a distinctive set of goals and values.

The Act creates responsibilities for 48 public bodies across Wales, with a commissioner to oversee them. It requires them to think and act for the long term, to integrate their activities with other public bodies, to involve citizens and communities in planning and delivery, to collaborate with communities and other public bodies, and to focus on preventing public ills. Businesses are expected to deliver prosperity and growth in ways that protect the environment, create good jobs and generate wealth for Welsh communities. Many of the deliverables associated with Act are concerned with the relationship between government and business, including training, business support and procurement. In other words, part of the work of the Act is to communicate its values to Welsh businesses.

3. Methodology

We found and created two sets of website text samples: (1) a reference set of texts relating to the WFGA and (2) the text from a structured sample of 5,431 Welsh companies 'website. Text embeddings were used to measure how closely different pieces of text resemble each other, making a comparison between the language used and frequency and saliency of key terms. The reference set of texts comprised 26,489 words that explain and represent the values and principles underpinning the legislation, drawn from the extant Welsh Government literature about the act and its implementation. We deployed text vectorisation to create a model for measuring the semantic similarity of large samples of unstructured text. This was subjected to a normalising process, by which we were also able to find co-incident topics that arise in the wild alongside the stated values of the WFGA. This repository was used to train the model.

We then created a sample of website texts to test against the model. We maintain a structured sample of 5,431 Welsh companies selected to represent the distribution of businesses by SIC code in each of Wales '22 Unitary authorities, providing a highly representative picture of the make-up of the Welsh economy that has previously been used for a national survey (Hogan et al., 2023). We crawled these websites to a depth of 0 (only the front page), to retrieve January 2024 text. The model analysed the prevalence and salience of topics related to the WFGA, which would become a Welsh Future Generations (WFG) Score for each company. Text was extracted only from HTML found on the websites; any text found between the script and styling tags was included. At the time of collecting, 272 of the 5,431 in the sample were inaccessible.

We generate a WFG score by converting the target text and the reference text to vectors using Google's "Universal Sentence Encoder". The text of each as a whole was encoded, rather than

an aggregate of the sentence level. We then find the cosine distance between the vectors, giving us a score from -1 to 1 where -1 is as different from the reference texts as it's possible to be, and 1 is exactly the same as the reference texts. This method is based on two influential papers on Word2Vec (Mikolov et al 2013) and Universal Sentence Encoder (Cer et al. 2018). The application of these methods for semantic text representation has been recently reviewed (Worth 2023), and others have developed it for similar purposes of textual semantic comparison (Agarwala et al. 2021; Jangabylova, A. et al. 2022). There are precedents for using it to measure shifting values in public discourse that reflect policy change (Rodman 2020). The end product of our process is a metric, a number between 0 and 1, for each company that reflects the degree of similarly between the material produced by an individual organisation and the content of the reference WFGA-encoded set. In other words, this offers a measure of the extent and degree to which the content of a business's website reflects or embodies the span of issues represented in discussions of the legislation.

Our proposition is that a higher WFG score indicates that this business has engaged with the values expressed in the WFGA legislation, and has made an attempt to embody these values in its public communications. We hypothesise that the presence of text on a website that reflects the language and values of the Act may be used as a proxy measure for compliance with the Act. We believe this leap is possible because (a) the Welsh government is very clear about its intention, and clear that they want businesses to very closely replicate the Act in the language they use; (b) they are very prescriptive about what businesses should say; and (c) the Act and these values are very well documented by the Welsh government, with a large set of reference texts to train a model on. This leads us to have confidence that when we find a close text match on a business's website, it can be interpreted as an expression of the Act's values.

4. Results

The results below indicate that our method is capable of discriminating meaningful differences in the prevalence and contextual salience of values associated with the WFGA across regions, business sector and participants in a government innovation support program. To the extent that the WFGA represents an attempt to intervene in the values and priorities shaping the Welsh economy, the WFG scores of Welsh businesses provide a good overall indication of the Act's success in achieving its goals.

4.1. The WFG Score results against the national sample

The WFG Score for the national sample reflects a normal distribution, with most scores clustered between 0.3 and 0.4 (shown in Figure 1, below). This provides reasonable evidence that the metric is measuring a real variation in the digital presence of the target companies. It is to be expected that that distribution is shifted to the left, because a) it is highly unlikely that any

website would yield a perfect score, and b) the legislation is still relatively new, so it is still early in the process of educating the business community.

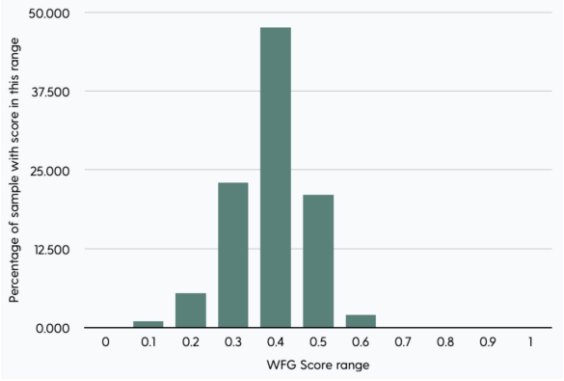


Figure 1. Distribution of WFG Scores for our national sample of Welsh companies.

4.2. The WFG Score results by Unitary Authority

The score for each Unitary Authority (UA) is averaged, to accommodate the fact that some UAs have many more businesses than others. Results (shown in Figure 2, below) indicate that the WFGA metric successfully distinguishes between UAs in Wales, with UAs representing the urban centres of Cardiff and the South scoring highest. This is to be expected, due to high business density, high levels of digitalisation and a high percentage of businesses selling to the public sector given their proximity to the seat of government.

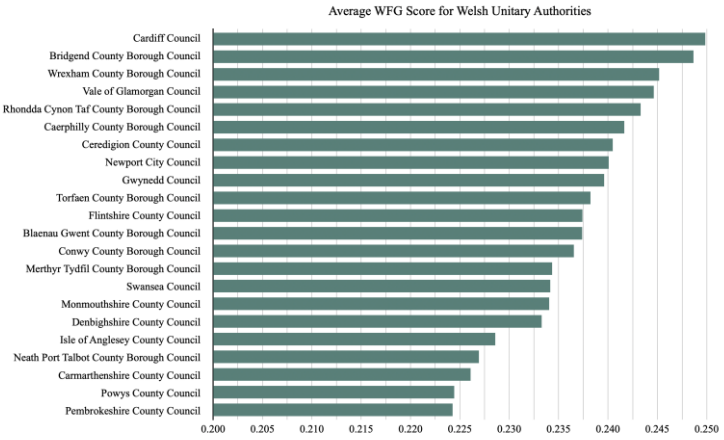


Figure 2. Average WFG Scores for Welsh Unitary Authorities.

4.2. The WFG Score results by business type

Our SIC group results (shown in Figure 3, below) also suggest that the WFGA metric successfully distinguishes between business sectors, with Business and IT (tending to be concentrated in the South and other urban areas) and Construction (tending to be engaged with government procurement and subject to strict criteria) scoring higher, while Retail/Wholesale Trade and Accommodation and Food Services are lower, as our graph below shows.

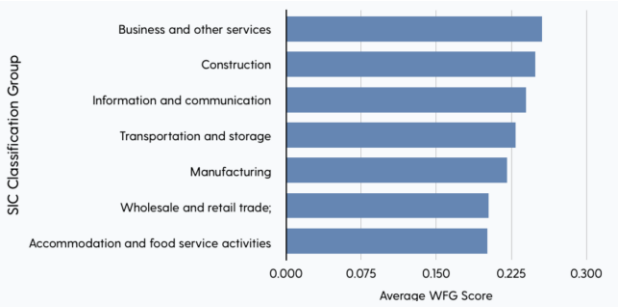


Figure 3. Average WFG Score by SIC Group.

We speculate that this difference may in part be due to the ways in which different kinds of business are required to comply with the Act. There are obvious places in which Construction businesses, for example, must comply, such as having environment policies on their websites, which is not the case for restaurants or holiday accommodation rentals. The highly digitalised kinds of services offered in the Business and IT sector trade significantly on credibility generated through their own digital properties, which is not true for many Accommodation and Food Services, who often use proprietary platforms such as AirBnB, Booking.com and Deliveroo to generate business. They do not therefore regularly update their websites, and so also miss the incidental and unconscious updates other sectors make regularly, that reflect more widespread changes in cultural values and attitudes promoted by the Act. This is substantiated by these sectors 'lower Digital Growth and Digital Maturity recorded in the Wales Digital Maturity Survey (Welsh Digital Maturity Survey 2023).

4.3. The WFG Score results for a government-funded innovation support programme

Our results for a sample of 340 companies that took part in a government-funded innovation support programme (shown in Figure 5, below) are shifted to the right, meaning they show systematically higher scores than the national sample. This suggests that direct contact with the government tends to lead to higher WFG scores, whether through selection or training.

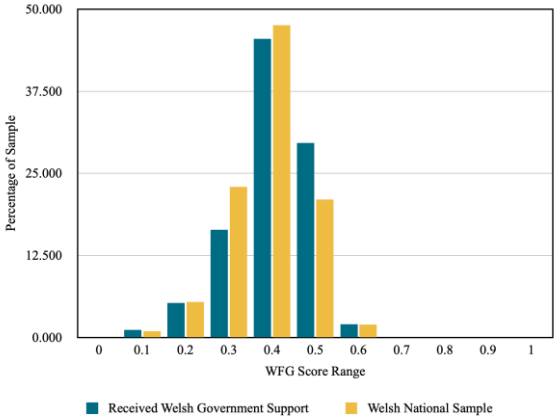


Figure 4. WFG Score comparisons.

4.4. English Counterfactual WFG Scores

We ran the WFG metric on an English counterfactual sample, where we matched English companies to the Welsh sample by sector (SIC code), company age, size (by accounts), and rural or urban (from their ONS postcode classification). As the graph below shows (Figure 5), the counterfactual sample performs considerably less well than the Welsh sample, as expected.

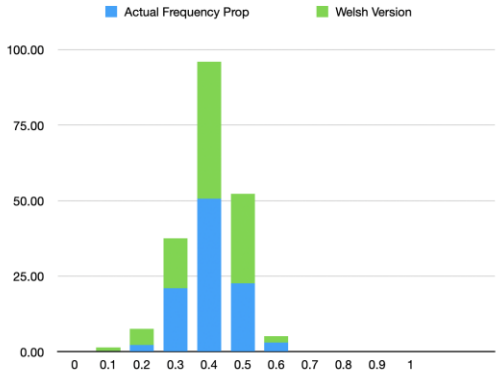


Figure 5. WFG Score of English counterfactual against the Welsh sample

5. Discussion

The results demonstrate the validity of the method, suggesting it is possible to use WebAI and word vectorisation to measure the prevalence of a set of ideas or values among a community of businesses, to measure differences between business type and geographic location, and to measure the impact of targeted government support programmes. We argue that it is possible to

make a meaningful comparison in this case between the reference texts and the target texts because of the volume of documentation and its prescriptiveness. Possible future applications of this include monitoring the effects of the act over time as well as monitoring and evaluating the results of specific interventions.

We have applied a similar method to developing other metrics, such as our Sense of Place metric, developed for a funding evaluation programme (Turner et al. 2024). We are currently working on deploying a DEI (diversity, equity and inclusion) metric to evaluate the DEI values of nursing colleges in the US; given the recent legal challenges and rapid change around DEI in US higher education (Murray et al. 2023) this is a timely task, offering the ability to monitor changing responses as they occur.

5.1. Limitations and drawbacks

This method relies on businesses having a website and the web domain being discoverable and validated. It works well for SMEs and large corporate entities, but may miss sole traders and micro businesses that do not have websites, that rely on word of mouth or social media platforms to do business.

5.2. Where next

Our next steps are to test the results against convergent validities, including testing against an English counterfactual sample and historic website data. We have a method for collecting website data for our sample over using the Wayback machine, and we know the date the legislation took effect (2015), so it should be possible to find out if there is a noticeable change in average scores before and after the Act's implementation.

Having established the premise for measuring the proliferation of values by regulatory bodies in the context of the WFGA, which represents a fairly well-defined sandbox, we would like to try applying it to other policy areas such as DEI, financial regulation and EU AI regulation.

References

Agarwala, S. et al. (2021). Detecting Semantic Similarity Of Documents Using Natural Language Processing. *Procedia Computer Science*, Vol 189, 128-135, ISSN 1877-0509, doi: 10.1016/j.procs.2021.05.076.

Cer, D. et al. (2018). Universal Sentence Encoder. arXiv:1803.11175v2 **[cs.CL]**, https://doi.org/10.48550/arXiv.1803.11175.

Future Generations Commissioner for Wales. Retrieved from https://www.futuregenerations.wales/about-us/future-generations-act.

Jangabylova, A. et al. (2022). Greedy Texts Similarity Mapping. Computation, 10, 200. https://doi.org/10.3390/computation10110200.

Mikolov, T. et al. (2013) Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781v3 **[cs.CL]**, https://doi.org/10.48550/arXiv.1301.3781.

Murray, T.A. et al. (2023). Anti-DEI legislation targeting colleges and universities: Its potential impacts on nursing education and the pursuit of health equity. Nursing Outlook 71 (2023), https://doi.org/10.1016/j.outlook.2023.101994.

Rodman, E. (2020). A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors. *Political Analysis*, *28*(1), 87–111. doi:10.1017/pan.2019.23.

Turner, D., Ailes, O., Thomas, M., Hogan, A. "Welsh Government SMART Suite Evaluation". Cardiff, Welsh Government, Wavehill Limited. Interim 2024.

United Nations (2021). Our Common Agenda: UN policy brief. Retrieved from https://www.un.org/sites/un2.un.org/files/our-common-agenda-policy-briefs-a-quick-summary.pdf

Welsh digital maturity survey 2023. Retrieved from https://welshdigitalmaturitysurvey2023.eticlab.co.uk/.

Welsh Government (2022). Well-being of Wales: 2021. Retrieved from https://www.gov.wales/wellbeing-wales-2021-html.

Worth, P. (2023) Word Embeddings and Semantic Spaces in Natural Language Processing. *International Journal of Intelligence Science*, **13**, 1-21. doi: 10.4236/ijis.2023.131001.

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17830

Augmenting the Italian Third Sector registry using non-profit organisations' websites

Carlo Bottai[1], **Francesco Trentini**[2,3,4], **Anna Velyka**[2]

[1]Department of Economics, Management and Statistics (DEMS), University of Milano-Bicocca, Italy, [2]Department of Statistics and Quantitative Methods (DiSMeQ), University of Milano-Bicocca, Italy, [3]Interuniversity Research Center for Public Services (CRISP), [4]Laboratorio "R. Revelli", Italy.

How to cite: Bottai, C.; Trentini, F.; Velyka, A. 2024. Augmenting the Italian Third Sector registry using non-profit organisations' websites. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17830

Abstract

This paper presents a framework for enriching and complementing administrative data from the Italian Third Sector Single National Register (RUNTS) with textual content extracted from the websites of the non-profit organisations listed in it. Through an automated web-scraping process we associate a website to each organisation and extract from its textual content information to describe the areas of the entity's actual economic activity. We develop a machine learning classifier to allocate each organisation into standardised categories of the International Classification of Non-profit Organisations. We further explore collected web data to identify other dimensions of non-profit operations. Enriching administrative registers with web data can yield trustworthy and detailed insights into the landscape of non-profit economic activities. Obtained results open up opportunities for further research of the labour market and economic development generated by the Third Sector, as well as comparative analysis with the sector of for-profit enterprises.

Keywords: *Third Sector; Administrative data enrichment; Big Data; Web scraping; NLP.*

1. Introduction

The non-profit sector contributes largely to the Italian economy. According to the latest figures, in 2021 more than 363,499 non-profit entities were active in the country, employing 870 thousand workers directly and activating 4.7 million volunteers, vis-à-vis 332,266 active enterprises with more than five employees, almost 23 million employees and 59 million inhabitants in the same period (Italian National Statistical Institute, 2023). They operate in a variety of economic sectors, ranging from social assistance and healthcare to social cohesion, sports and cultural activities.

Nonprofit entities are of great interest since they operate following market principles and an efficient use of production factors, while not pursuing profit. These enterprises embrace a community-oriented model, which aims at keeping the surplus within the entities that generated it, fostering both investment and improvement in social and economic conditions in the context in which they are embedded (Evers & Laville, 2004).

In Italy, the discipline of the non-profit sector has been reviewed in 2017, with the Legislative Decree 117/2017, known as Codice del Terzo Settore (hereon, the Code). The Code institutes seven main categories of Third Sector Entities (Enti del Terzo Settore, ETS hereon): volunteer organisations, social promotion associations, social enterprises, philanthropic entities, networks of entities, mutual aid societies, and other entities. The Code enumerates the set of activities of general interest that every ETS must primarily pursue. The qualification as ETS introduces requirements in terms of governance and transparency of their accounts; at the same time, they benefit from special tax exemptions, they can sign partnerships with public entities and receive public funds. Among the most important innovations introduced by the Code is the establishment of a national electronic register of ETS, the Registro Unico Nazionale del Terzo Settore (RUNTS), owned by the Ministry of Labour and Social Policies, that substituted more than 350 local registries. An entity listed in the RUNTS gains the ETS title and acquires rights and obligations described by the law. The RUNTS is publicly available through a web interface (https://servizi.lavoro.gov.it/runts/) and includes a long list of characteristics of each ETS.

Our work leverages this new administrative source to establish a framework for data enrichment using entities' websites. In recent years a similar framework has been developed and widely applied to for-profit enterprises (see e.g., Daas & van der Doef, 2020; Kinne & Axenbeck, 2020; Crosato et al., 2023; Bottai et al., 2022; Abbasiharofteh et al. 2023). Indeed, nowadays firms commonly use corporate websites to reach customers with information about their products and business activities. Therefore, it is possible to infer the economic activity of a firm by looking at the content of its corporate websites, at least to some extent (Domènech et al., 2012). Non-profit organisations as well largely rely on their own website to communicate with their stakeholders. Consequently, these websites can be leveraged as a—easily accessible and constantly updated—information source to enrich the—already available and trustable—administrative sources. However, such kinds of unstructured sources require some careful clenching and pre-processing to mine useful and meaningful information from them (see e.g., Daas et al. 2015; Rammer & Es-Sadki, 2023). The rest of this work describes this process in detail. Then, some preliminary results and future developments are discussed.

2. Data and methods

The RUNTS (the register, hereafter) is operated by the Italian Ministry of Labour and Social Policies (MLPS).

2.1. The master database: the RUNTS

The register contains information about 123,221 ETSs, seventy-five per cent of which are either *associazioni di promozione sociale* (APS) or *organizzazioni di volontariato* (ODV), both voluntary organisations—the former carrying out activities primarily for the benefit of third parties, while the activities of the latter may be primarily for the benefit of their own members. For each entity, we know its tax code (*codice fiscale*), business name, legal type, headquarter's municipality, the date of enrolment in the register, and the generalities of the legal representative.

2.2. External data collection

The aim of this work is to enrich the information available on the RUNTS by exploiting textual content extracted from the ETSs' websites. Therefore, we start collecting the business name, tax code, and legal form of each entity as recorded in the register. After some minimal cleaning pre-processing of these strings, we search on Microsoft's Bing web search engine for queries combining these information pieces—like *("amici degli animali" OR 97324990385) APS*—, in a similar fashion to van Delden et al. (2019). For each ETS, we collect up to ten results returned by the search engine. We extract and preserve only the web domain of each result and we exclude those that are linked to multiple ETSs. Then, for each ETS we loop over the remaining web domains, and we scrape a few pages of this domain hoping to find at least one information piece that can verify the matching between the web domain detected and the ETS under scrutiny; see Barcaroli et al. (2016) and Bottai et al. (2022) for a similar approach. Specifically, we search on the home page of the website as well as on pages like "about us" or "contacts", and we consider as positively assigned these websites naming at least one among the tax code, headquarter's county (*provincia*), or full name of the legal person of the ETS linked with that website. While looping over the domains retrieved on Bing, we give priority to those with high textual similarity to the ETS's business name and to these high in the search engine's results. We stop the loop at the first positive match.[1]

To assess the quality of this procedure, we extracted 119 ETSs at random, representative of the population about the geographical area (NUTS 2) and legal type (APS, ODV, social enterprise, etc.). By searching on Bing for this list of entities, we detected at least a web domain for 103 of them (86.6%). Of these, we selected the first result returned by the search engine (we plan to extend this step to include all the first ten results that we got from the search engine) and we scraped a few pages of the website. The scraper has been able to retrieve the website of 81 ETSs (78.6%). Of each of these, we further verified the accuracy of the matching between the ETS in question and the website detected. We developed an automatic classifier that classified each of

[1] The authors are available to provide further information details about these steps.

the 81 cases as either a positive or negative matching. The classifier returned 35 positive and 46 negative matchings. We manually inspected the same matches and verified 31 true positives and 39 true negatives. Therefore, the classifier shows high *sensitivity* (81.6%) and *specificity* (90.7%).

Even though further development is needed, we believe that these preliminary results are encouraging. Most of the improvement must be obtained by improving the quality of the results obtained from the search engine. We believe that, by extending the number of results considered from one to ten (as already mentioned), we will be able to obtain a significant improvement in this respect.

2.3. Classification of economic activity

The RUNTS provides information concerning the main fields of activity of each ETS, according to the International Classification of Non-Profit Organisations (ICNPO) and Article 5 of Legislative Decree 117/2017. Nonetheless, the information is self-declared and not compulsory. Self-declaration may induce losses in precision, as ETSs may declare a smaller number of areas of activity, understating the actual scope of their activities, or be led to declare only those activities that are closer to their statutory mission. For this reason, ETS' web pages are a rich source of information on the actual activities put in place by the ETSs.

A number of classification systems exist at national and international levels to categorise non-profit sector organisations depending on the scope of their activities. At the national level, classifications often coincide with government frameworks and classify ETS based on their primary purpose and economic activities (like the NACE in the EU). The United Nations has developed the International Classification of Non-Profit Organisations (ICNPO) that provides standards and taxonomies specifically for non-profit organisations based on their activities, primary purpose, and mission.

For statistical accounting, the Italian National Institute of Statistics categorises ETS operating in Italy according to ATECO (the Italian version of the NACE). Since the ATECO hasn't been developed specifically for the non-profit sector, it doesn't provide enough categories to account for all the possible range of activities of these organisations (12 macro categories of ICNPO correspond to only 5 macro categories of ATECO). Instead, the ICNPO classification has 12 macro and 39 subcategories. In addition to being recognised as meeting the OECD and UN international standards, this classification provides a brief description of the organisations that fall under each subcategory (United Nations, 2018).

Several studies in the past have investigated the possibility of automatically identifying an appropriate category of non-governmental organisations. These studies used both international classifications (e.g., for identifying appropriate ICNPO category for 5,000 Austrian non-profit organisations; see Litofcenko et al., 2020) and national classifications (e.g., remapping the US nonprofit sector by reassigning multiple NTEE codes to organisations with purposes across

various domains; see Ma, 2021). The results of this research have been subsequently used to develop a national classification of charity organisations in the United Kingdom (UK Charitable Activity Tags) and an automated system for determining the appropriate category for 4,200 organisations (Damm & Kane, 2022).

Methods that were previously applied to analyse and classify textual data on ETS can be summarised into three categories: rule-based dictionary approach, supervised learning, and unsupervised learning. Data enrichment techniques proposed in this paper allow us to overcome the limitations of previous studies, providing more information about the ETSs activities. To deploy a trained classifier to automatically assign ICNPO category to new ETS based on the text information extracted from their websites we will apply Natural Language Processing (NLP) and test several supervised learning techniques in the following steps:

- **Pre-processing** of text data collected from websites of non-profit organisations: clean the text data by removing punctuation, stopwords, and HTML tags; tokenise the text into individual words or phrases.
- **Feature Extraction** from the pre-processed data: convert text into numerical features to be used by machine learning algorithms: bag-of-words and word embedding.
- **Classification Category Description Parsing**: parse the short descriptions of each ICNPO category to extract keywords that are indicative of the category's focus area.
- **Model Training**: the information about ICNPO classification available in RUNTS for some ETSs will serve as a basis for training the model. We will train a classifier to map the numerical representations of textual data to predefined ICNPO categories based on labelled training data (Naive Bayes, Support Vector Machines, Decision Trees, and Convolutional Neural Networks, Bidirectional Encoder Representations from Transformers).
- **Model Evaluation**: some ETS already classified into one of the ICNPO classes will be used as a test set to evaluate the models' performance. A stratified split will help to ensure the proportional representation of the data. Consequently, several metrics will be calculated to demonstrate the effectiveness of the suggested classification model. The accuracy metric will reveal the overall model's performance by denoting the proportion of correctly classified ETS out of the total number of ETS from the test set. Since we expect that not all the ICNPO classes are equally represented, we will also employ other performance metrics. Precision will demonstrate the accuracy of the model in correctly assigning ICNPO classes to companies based on their web text data. Recall metric will indicate the model's ability to capture all ETS of a particular ICNPO class. Finally, a harmonic mean of precision and recall metric (F-1 score) will denote the overall reliability of our model.

2.4. Classification of other dimensions

The information concerning the sectors of activity in which each ETS is involved is the primary dimension of interest to integrate the information available through the administrative register.

In fact, other dimensions can be approached using website data. Other dimensions are the beneficiaries of their activities, the type of projects developed by each ETS, the stakeholders whom they are in relation with, and the funding they receive. We are testing the use of the websites to address these dimensions and extract useful information. The use of web data has the additional advantage of storing pages to run offline analysis, so that they can be queried in subsequent moments over dimensions that were not foreseen in the initial stages of the work.

3. Discussion and future work

The use of web data for the production of both fresh statistics or for the enrichment of administrative sources has reached a stage of maturity and applied in a large set of cases, including the production of official statistics, introducing changes in the production process, in term of theoretical frameworks and required competences (Pfeffermann, 2015; Ricciato, Wirthmann & Hahn, 2020). It is especially important to consider the data generating process of web data is complex. Agents produce data with a multitude of schemes and formats, usually not standardised. Moreover, the processes generating these data are variable, too. Therefore, a lot of attention is needed to understand the veracity of these data, that is, which phenomena these data represent. The opportunity presented by these unique data characteristics is associated with the central challenge of interpreting the content of the recorded data and the correct way to treat and process them. Administrative registers are therefore a valuable starting point, given that they provide the universe of observations concerning specific social facts, generated by law provisions. The RUNTS collects the universe of Third Sector Entities in Italy and enriching it by means of web data can provide trustable, timely and fine-grained information on their operations. The ETSs are of great interest given that they provide essential services to a large part of the population and complement publicly provided services with a diverse business model.

Having a timely and precise knowledge of the ETSs' area of activity and types of interventions is pivotal to coordinate interventions in areas of general interest through partnerships or agreements with the public administration, which is also one of the pillars of the 2017 reform. It also allows us to analyse the social capital of local communities and study in depth the relationship of this last with economic development. Furthermore, the abundance of individual information in the RUNTS, such as the tax code and the names of the legal representatives, makes the database virtually linkable to a huge variety of data sources. Of particular interest are the enterprise characteristics and performance, on the business demography side, and, on the labour market side, the volume and quality of employment generated by the Third Sector. Possible extensions would also include the analysis of the hyperlinks connecting the ETSs' websites to each other (*webometrics*), e.g., to analyse the *networks of ETSs*, a kind of ETS regulated by the RUNTS, or to study the business relationship among ETSs (see e.g., Vaughan et al. 2006 and Abbasiharofteh et al. 2023). At the same time, it allows running a comparative

analysis with the Second Sector, i.e., for-profit enterprises, in terms of generated employment and their business demography and potentially filling a gap in the national statistics, that have a hard time at properly representing such an informal sector by only relying on conventional information sources.

Acknowledgements

Carlo Bottai and Francesco Trentini thank the University of Milano-Bicocca for sponsoring this work under the 'Bicocca Starting Grants 2023–2024' funding schema.

References

Abbasiharofteh, M., Kinne, J., & Krüger, M. (2023). Leveraging the digital layer: the strength of weak and strong ties in bridging geographic and cognitive distances. *Journal of Economic Geography*, lbad037. doi:10.1093/jeg/lbad037

Barcaroli, G., Scannapieco, M., & Summa, D. (2016). On the use of Internet as a data source for official statistics: a strategy for identifying enterprises on the Web. *Rivista Italiana di Economia, Demografia e Statistica*, 70(4), 25–41.

Bottai, C., Crosato, L., Domènech, J., Guerzoni, M., & Liberati, C. (2022). Unconventional data for policy: Using Big Data for detecting Italian innovative SMEs. In *Proceedings of the 2022 ACM Conference on Information Technology for Social Good* (pp. 338–344). New York, NY: Association for Computing Machinery. doi:10.1145/3524458.3547246.

Crosato, L., Domènech, J., Liberati, C. (2023). Websites' data: a new asset for enhancing credit risk modeling. *Annals of Operations Research*, 1–16. doi:10.1007/s10479-023-05306-5

Daas, P.J.H., & van der Doef, S. (2020). Detecting innovative companies via their website. *Statistical Journal of the IAOS*, 36(4), 1239–1251. doi:10.3233/SJI-200627

Daas, P.J.H., Puts, M.J., Buelens, B., & van den Hurk, P.A.M. (2015). Big Data as a Source for Official Statistics. *Journal of Official Statistics*, 31(2), 249–262. doi:10.1515/jos-2015-0016

Damm, C., & Kane, D. (2022). Classifying UK charities' activities by charitable cause: a new classification system. *Voluntary Sector Review*, 1–27.

van Delden, A., Windmeijer, D., ten Bosch, O. (2019). Searching for business websites. Discussion Paper. Statistics Netherlands (CBS).

Domènech, J., de la Ossa, B., Pont, A., Gil, J.A., Martinez, M., & Rubio, A. (2012). An Intelligent System for Retrieving Economic Information from Corporate Websites. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology* (vol. 1, pp. 573–578). Washington, DC: IEEE Computer Society. doi:10.1109/WI-IAT.2012.92

Evers, A., & Laville, J.-L. (Eds.) (2004). *The Third Sector in Europe*. Edward Elgar. doi:10.4337/9781843769774

Italian Ministry of Labour and Social Policies (2024). *Registro Unico Nazionale del Terzo Settore* [Data set]. Retrieved from https://servizi.lavoro.gov.it/runts/ on 2024-03-14.

Italian National Institute of Statistics (2023) Censimento Permanente delle Istituzioni Non Profit.

Kinne, J., & Axenbeck, J. (2020). Web mining for innovation ecosystem mapping: A framework and a large-scale pilot study. *Scientometrics*, 125(3), 2011–2041. doi:10.1007/s11192-020-03726-9

Litofcenko, J., Karner, D., & Maier, F. (2020). Methods for classifying nonprofit organizations according to their field of activity: A report on semi-automated methods based on text. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, 31(1), 227–237.

Ma, J. (2021). Automated coding using machine learning and remapping the US nonprofit sector: A guide and benchmark. *Nonprofit and Voluntary Sector Quarterly*, 50(3), 662–687.

Pfeffermann, D. (2015) Methodological Issues and Challenges in the Production of Official Statistics: 24th Annual Morris Hansen Lecture, *Journal of Survey Statistics and Methodology*, 3(4), 425–483. doi:10.1093/jssam/smv035

Rammer, C., & Es-Sadki, N. (2023). Using big data for generating firm-level innovation indicators - a literature review. *Technological Forecasting and Social Change*, 197, 122874. doi:10.1016/j.techfore.2023.122874

Ricciato F., Wirthmann A., & Hahn M. (2020) Trusted Smart Statistics: How new data will change official statistics. *Data & Policy*, 2, e7. doi:10.1017/dap.2020.7

United Nations (2018). *Satellite Account on Non-profit and Related Institutions and Volunteer Work* [Data set]. Retrieved from https://unstats.un.org/unsd/nationalaccount/docs/UN_TSE_HB_FNL_web.pdf

Vaughan L., Gao Y., & Kipp M. (2006) Why are Hyperlinks to Business Websites Created: A Content Analysis. *Scientometrics*, 67(2), 291–300. doi:10.1007/s11192-006-0100-6

Using texts to measure proximity between firms

Alessandro Marra[1,2], Cristiano Baldassari[2,3]

[1]Dipartimento di Economia, Università G. d'Annunzio di Chieti-Pescara, Pescara, Italy, [2]Explo, spinoff accademico, Università degli Studi G. d'Annunzio di Chieti-Pescara, Pescara, Italy, [3]Dipartimento di Neuroscienze, Università G. d'Annunzio di Chieti e Pescara, Italy.

How to cite: Marra, A.; Baldassari C. 2024. Using texts to measure proximity between firms. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. p. 148. https://doi.org/10.4995/CARMA2024.2024.19017

Abstract

Measuring the similarity (or proximity) between firms is an increasingly important endeavor in both academic research and policy-making circles. The proposed analysis employs text mining and semantic algorithms for processing of descriptive data. The adopted approach enables policymakers to survey the landscape and pinpoint the targets for policy interventions more accurately. Employing graph embedding techniques aids in identifying clusters of companies based on their specializations, technologies, competencies, and knowledges within a lowdimensional space. Through Role2Vec embedding, which interprets the network of connections between nodes and captures their structural relationships, we generate a vector representation for each node, facilitating classification, clustering, link prediction, and more. We refine this by mapping the multi-dimensional embedding to a two-dimensional space using t-distributed stochastic neighbor embedding (t-SNE), creating an intuitive visualization.

Keywords: *Business proximity; text data; industrial clusters.*

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17820

Evaluating coherence in AI-generated text

María Olmedilla[1], José Carlos Romero[2], Rocío Martínez-Torres[3], Nicolas R. Galvan[4], Sergio Toral[4]

[1]SKEMA Business School, Université Côte d'Azur, France, [2]Applied Computational Social Sciences Data-Intensive Governance-Institute, Université Paris Dauphine-PSL, France, [3]Facultad de Ciencias Económicas y Empresariales, University of Seville, Spain, [4]E. T. S. Ingeniería, University of Seville, Spain.

How to cite: Olmedilla, M.; Romero, J. C.; Martínez-Torres, R.; Galvan, N.; Toral, S. 2024. Evaluating coherence in AI-generated text. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17820

Abstract

This study examines the role of coherence in AI-generated online reviews and its effect on perceived authenticity and consumer trust. By applying advanced metrics like BERT Score, BART Score, and Disco Score, the research analyzes the coherence of AI-generated text using Generative AI models, specifically Llama-2, on Amazon beauty product reviews. Results indicate that AI-generated reviews exhibit higher coherence compared to human-generated content, suggesting that Generative AI can produce seemingly authentic content. This finding challenges the ability to distinguish between human and AI-generated reviews, raising important questions about consumer trust in digital marketplaces. The study underscores the importance of coherence in online content's credibility and opens avenues for further research on Generative AI's role in e-commerce.

Keywords: *Generative-AI; Online reviews; Llama-2, BERT, Coherence.*

1. Introduction

In the digital era, the authenticity and integrity of online content have become paramount, especially with the proliferation of user-generated and AI-generated texts in digital marketplaces. The concept of coherence in online reviews, which is a significant component of texts, not only signifies authenticity but also impacts consumer trust and decision-making. This paper aims to delve into the intricate relationship between coherence and perceived authenticity in AI-generated text, highlighting the important role of advanced computational and linguistic tools in this evaluation process.

There has been some efforts among researchers to underscored the critical role of advanced computational tools and linguistic theories to explore the coherence in AI-generated text. For

instance, Ai et al. (2019) discuss the integration of text coherence in text generation, emphasizing the significance of coherence metrics like semantics and syntax-based coherence in enhancing text generation. Likewise, Elkhatat et al. (2023) investigate the distinction between human and AI-authored content. Their work reveals the capabilities and limitations of AI content detection tools, highlighting the ongoing challenge in accurately identifying AI-generated text.

Furthermore, the advent of sophisticated computational linguistic tools, such as BERT Score, BART Score, and Disco Score, has revolutionized our ability to assess coherence in AI-generated text, their effectiveness in maintaining the integrity of online reviews, and the implications for consumer trust in an era increasingly dominated by generative AI.

2. Research Background

2.1. Generative AI and online reviews

The exploration of Generative Artificial Intelligence (AI) in the context of online reviews represents a growing field that holds promise for reshaping e-commerce and user interaction on online platforms. Generative AI, leveraging models such as Transformers, has extended its utility beyond traditional applications like image processing and natural language processing to the generation of online reviews (Bandyopadhyay et al., 2023). The exploration of generative AI in online reviews seeks to understand its impact on credibility, authenticity, and usefulness (Hu et al., 2012). Besides, the effectiveness of generative AI in producing online reviews that are perceived as helpful and authentic by users is crucial for its application in e-commerce platforms, a critical factor for consumer trust and decision-making

Nevertheless, the integration of generative AI into the creation of online reviews is not without its challenges. Studies contrasting prior laboratory evidence have shown that the use of generative AI in text generation can result in a decline in the quality of online reviews. This effect is particularly pronounced among non-expert reviewers, though it also leads to an increase in the quantity of content produced per reviewer, highlighting a trade-off between quality and volume (Knight & Bart, 2023). In the case of the recommender systems, generative AI has been employed to enhance the informativeness and efficiency of user-generated reviews. By combining traditional collaborative filtering methods with advanced deep learning architectures for text processing, researchers have achieved superior performance in recommendation accuracy compared to baseline systems. This demonstrates generative AI's potential to significantly improve the personalization and relevance of online recommendations (Shalom et al., 2019). Another critical application of generative AI in online reviews is in spam detection. Innovative methods based on aspect-level analysis of reviews have shown promise in identifying and mitigating spam content, thereby protecting the integrity of online review

platforms (Wang et al., 2022). Generative AI's capability for content generation has also been explored in review generation tasks, in thir regard the research by Zang and Wan (2017) focused on long review generation within the encoder-decoder neural network framework. This line of work addressed the challenges of automating review generation to produce helpful and persuasive online content. Furthermore, online reviews are central in influencing consumer buying choices, although only a small number of users dedicate effort to crafting constructive reviews. Fortunately, the latest advances in deep neural networks presents a promising avenue for generating content that closely resembles authentic reviews (Kaghazgaran et al., 2020).

In summary, the intersection of generative AI with online reviews encompasses a diverse range of applications, from enhancing the personalization of recommendations to improving the integrity and usefulness of user-generated content. Despite the challenges related to content quality, the ongoing research and development in this area (Ooi et al., 2023) underscore generative AI's transformative potential in e-commerce and beyond.

2.2. Measuring the coherence between true and fake reviews

The foundational principle that coherence is indispensable for effective written discourse, as argued by Bamberg (1983), sets the basis for understanding its relevance in evaluating online reviews. Giora (1985) extends this by emphasizing the role of "aboutness" or the discourse topic in achieving text coherence, thus highlighting the intrinsic connection between coherence and the pragmatic formation of text. This theoretical framework can help assessing the credibility of online reviews, where coherence may indicate authenticity.

Furthermore, the methodology developed by Foltz et a.l (1998) for measuring textual coherence with latent semantic analysis introduces a quantitative approach to this qualitative attribute, suggesting that coherence can be systematically analyzed and assessed. Likewise, Cui et al. (2017) contribute to this domain by employing deep learning models to evaluate text coherence, illustrating the potential of advanced computational techniques in discerning well-organized from poorly structured texts, a distinction crucial for identifying fake reviews.

Additionally, the role of prior knowledge in text comprehension introduces a nuanced perspective on how individual differences in knowledge coherence could influence the perception and evaluation of online reviews (McCarthy & McNamara, 2021). This aspect is particularly relevant in the context of spam detection, such as the approach proposed by Yang (2015), which uses coherence metrics to distinguish between genuine and spam reviews.

In this regard, Singh et al. (2020) empirically demonstrate that fake news articles show lower textual coherence compared to legitimate counterparts, an insight that can be extended to online reviews, suggesting that incoherence may be a sign of fraudulent content. This is validated by Liu et al. (2024), who propose a coherence-based ranking system for online reviews,

highlighting the influence of coherence on consumer decision-making efficiency and its significance in enhancing marketplace integrity.

BERT and BART Scores have emerged as key in understanding the structural, stylistic, and semantic coherence of online reviews. Authors such as Koh (2011) have developed methodologies that quantify the sentiment in online reviews, acknowledging the profound impact of linguistic coherence on product sales and the importance of sentiments beyond mere numerical ratings. This has significant implications, as demonstrated by Purnawirawan et al. (2014), where coherence in review content alongside source credibility was shown to significantly influence readers' perceptions and intentions in online review scenarios. In evaluating Polish texts' coherence, Telenyk et al. (2021) employed neural networks and a pre-trained BERT model, showcasing the evolving methodologies in coherence assessment.

The Disco Score, although less prominent in the literature, also has potential in defining the fine line between authentic and fraudulent content. Research by Bhāle and Tongare (2018) focused on profiling online hotel reviews to distinguish between genuine and fake content, comparing different travel websites to understand the variances brought about by online review structures.

Consequently, all these research controbutions show the importance of coherence in online reviews, not only as a symbol of authenticity but also as a determinant of consumer trust and decision-making. By leveraging advanced computational tools and linguistic theories, researchers and practitioners alike can better navigate the complexities of online consumer feedback, ensuring the reliability and integrity of user-generated content in digital marketplaces.

3. Methodology

The datasets employed in this paper consist of online reviews across various Amazon product categories, which were obtained from the work of Ni et al. (2019). These datasets contain millions of reviews classified into 29 different product categories. To test our methodology we have focused only on the product category "*Beauty Products"*.

Limiting the analysis to one product category ensures more coherence among the reviews, as they concern the same topic. Our dataset has been preprocessed to keep only the following useful information: "review ID, text of the review and the *verified purchase* label. Such label allows us to verify the authenticity of the authorship of the review, so we can asure that the review has been written by a human verified by Amazon.

To develop the generation of artificial reviews (AI-generated text) we use a Large Language Model (LLM), particularly Llama-2, which is ranked currently as one of the state-of-the-art for open-source models and widely used in research (Touvron et al. 2023). We need to fine-tune the model to generate more accurate text regarding the topic analyzed.

Figure 1 shows an overview of the methodology. In *step #1* we fine-tune a Llama-2 model to generate artificial reviews. We use two different models, 7-billion-parameter and 13-billion-parameter, which differ in the size (number of parameters) of the model. Due to limitations in computational resources, it was not feasible to employ a larger model, such as the 70-billion-parameter one. As an input to the fine-tuning we give to both models 30,000 reviews from the category "*Beauty Products"* that have the *verified purchase* label. Subsequently, in *step #2*, we generate 10,000 artificial reviews using each model, aiming to compare these with verified real reviews. In the last step, step #3, we apply the three metrics most commonly used in the literature for measuring text coherence: *BERTScore*, *BARTScore*, and *DiscoScore*. These three metrics are applied to three distinct datasets: a chunk of real reviews obtained from the *verified purchase* input data, reviews generated by the Llama-2-7b model, and reviews generated using the Llama-2-13b model.

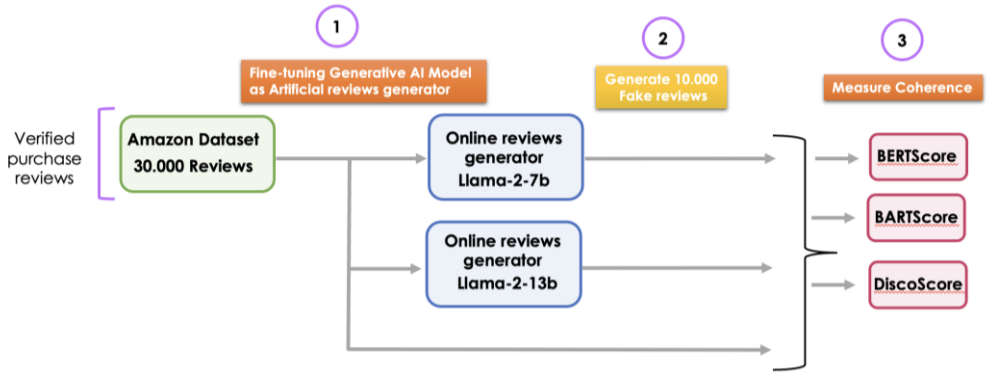


Figure 1. Outline of the methodology

BERTScore, *BARTScore*, and *DiscoScore* are three distinct metrics designed to evaluate text from various angles, emphasizing semantic similarity, text generation quality, and discourse coherence, respectively. *BERTScore* assesses semantic similarity by comparing the embeddings of tokens generated by BERT between a reference text and a candidate text. It calculates recall, precision, and F1 scores based on the maximum cosine similarities of token pairs, emphasizing the semantic overlap and accuracy of the information conveyed. *BARTScore*, on the other hand, leverages BART, a pre-trained text generation model, to predict words or sentences in the reference text, thus evaluating the generated text's ability to capture the intended meaning and information. It uses a log probability score, aiming for a score closer to zero, to gauge the effectiveness of text generation in terms of fluency and coherence. *DiscoScore* specifically targets discourse coherence using BERT to analyze how well sentences and ideas flow together in a text. It focuses on nominal and semantic entity focus, tracking important nouns and broader semantic entities to understand how ideas relate. This metric examines focus continuity and

coherence relationships, measuring the discourse coherence beyond mere semantic similarity or text generation quality.

4. Results

For fine-tuning the generative AI models, generating artificial reviews, and applying the targeted metrics, we used Google Colab's free membership, which offers a T4 GPU with 16GB of VRAM. We have systematically applied this metrics to 100 reviews on each dataset. We systematically applied these metrics to 100 reviews in each dataset. However, an extension of the analysis to a larger set of reviews was constrained by limitations in our computational resources.

Table 1 presents the preliminary results of our study, demonstrating consistency across the three metrics. *BERTScore* and *DiscoScore* report more coherence as the score approaches to 1, while *DiscoScore* reports more coherence with the score closer to 0. The results show that, contratiwise, AI-generated reviews show more coherence in all metrics: up to 10% more in *DiscoScore*, 30% in *BARTScore* and 3% in *BERTScore*. This can be explained as generative AI models are designed to ensure the coherence of the text they generate, in contrast to human users who might not prioritize coherence when writing reviews. Surprisingly, Llama-2-7b demonstrates greater coherence than Llama-2-13b, despite being a smaller model with apparently lower performance. This phenomenon could be attributed to the bigger LLM model requiring more refined fine-tuning, possibly involving a larger dataset of reviews or more optimized hyperparameters.

Table 1. Results of the three metrics and the 3 datasets analyzed

Metric Applied	***Verified Review***	***Llama-2-7b***	***Llama-2-13b***
BERTScore	0.842	0.867	0.862
BARTScore	-4.434	-3.131	-3.235
DiscoScore	0.780	0.860	0.849

5. Conclusions

In this work we have developed a methodology to analyze the coherence in AI-generated text. Through a process of fine-tuning we ensure that our generative AI model generates more accurate text regarding the topic analyzed, allowing a more objective comparising respect to the verified reviews. We use 30.000 verified amazon reviews to fine-tune two different generative AI models: Llama-2-7b and Llama-2-13b, which we use afterwards to generate 10,000 artificial reviews with each one. We apply systematically 3 state-of-the-art metrics (*BERT Score*, *BART Score*, and *Disco Score*) to measure the coherence through a subset of 100 reviews from the three datasets. Results show that AI-generated text shows more coherence, up to 30% more. These results highlight the challenge to distinguish between human and AI-generated reviews,

and the impact on consumer trust in digital marketplaces. In our current pipeline of work we plan to extend this analysis to a broader dataset of reviews, and to add more metrics to measure the quality of the text generated, as the readability or qualitative characteristics of the review (length, keywords, topic modelling, etc).

References

Ai, L., Gao, B., Zheng, J., & Gao, M. (2019, December). On Improving Text Generation Via Integrating Text Coherence. In *2019 IEEE 6th International Conference on Cloud Computing and Intelligence Systems* (CCIS) (pp. 6-10). IEEE.

Bamberg, B. (1983). What makes a text coherent?. *College Composition and Communication*, 34(4), 417-429.

Bandyopadhyay T., Saha S., Pal D., (2023). Beyond Imitation: Exploring Novelty in Generative AI, *International Journal of Advanced Research in Science Communication and Technology*, 3 (2).

Bhāle, S., & Tongare, K. (2018). An empirical investigation of gist helpfulness in online reviews. *Journal of Business and Retail Management Research*, 13(02).

Cui, B., Li, Y., Zhang, Y., & Zhang, Z. (2017, November). Text coherence analysis based on deep neural network. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 2027-2030).

Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, 19(1), 17.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2-3), 285-307.

Giora, R. (1985). Notes towards a theory of text coherence. *Poetics Today*, 6(4), 699-715.

Hu, N., Bose, I., Koh, N. S., & Liu, L. (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision Support Systems*, 52(3), 674-684.

Kaghazgaran, P., Wang, J., Huang, R., & Caverlee, J. (2020, July). Adore: Aspect dependent online review labeling for review generation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1021-1030).

Knight, S., & Bart, Y. (2023). Generative AI and User-Generated Content: Evidence from Online Reviews. Available at *SSRN 4621982*.

Koh, N. S. (2011). The valuation of user-generated content: a structural, stylistic and semantic analysis of online reviews. *Singapore Management University*.

Liu, Y., Qiao, D., & Li, X. (2024). In Coherence We Trust: Analyzing Effects of Discourse Coherence in Online Reviews. Available at *SSRN* 4714241.

McCarthy, K. S., & McNamara, D. S. (2021). The multidimensional knowledge in text comprehension framework. *Educational Psychologist*, 56(3), 196-214.

Ni, J., Li, J., & McAuley, J. (2019, November). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In Proceedings of the 2019 Conference on

Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 188-197).

Ooi, K. B., Tan, G. W. H., Al-Emran, M., Al-Sharafi, M. A., Capatina, A., Chakraborty, A., & Wong, L. W. (2023). The potential of generative artificial intelligence across disciplines: Perspectives and future directions. *Journal of Computer Information Systems*, 1-32.

Purnawirawan, N., Dens, N., & De Pelsmacker, P. (2014). Expert reviewers beware! The effects of review set balance, review source and review content on consumer responses to online reviews. *Journal of Electronic Commerce Research*, 15(3), 162-178.

Shalom, O. S., Uziel, G., & Kantor, A. (2019, September). A generative model for review-based recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems* (pp. 353-357).

Singh, I., Deepak, P., & Anoop, K. (2020). On the coherence of fake news articles. In *ECML PKDD 2020 Workshops: Workshops of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2020): Ghent*, Belgium, September 14–18, 2020, Proceedings (pp. 591-607). Springer International Publishing.

Telenyk, S., Pogorilyy, S., & Kramov, A. (2021). Evaluation of the coherence of Polish texts using neural network models. *Applied Sciences*, 11(7), 3210.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Wang, S., Jiang, W., & Chen, S. (2022, December). An Aspect-Based Semi-supervised Generative Model for Online Review Spam Detection. In *International Conference on Ubiquitous Security* (pp. 207-219). Singapore: Springer Nature Singapore.

Yang, X. (2015, January). One methodology for spam review detection based on review coherence metrics. In *Proceedings of 2015 international conference on intelligent computing and internet of things* (pp. 99-102). IEEE.

Zang, H., & Wan, X. (2017, September). Towards automatic generation of product reviews from aspect-sentiment scores. In *Proceedings of the 10th International Conference on Natural Language Generation* (pp. 168-177).

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17818

A Comparative Analysis of Companies Missing from the SABI Database through BORME Gazette Web Scraping

Xin-Hui Huang[1], Josep Domenech[2]

[1]ETSINF, Universitat Politècnica de València, Spain, [2]DECS, Universitat Politècnica de València, Spain.

How to cite: Huang, X.; Domenech, J. 2024. A Comparative Analysis of Companies Missing from the SABI Database through BORME Gazette Web Scraping. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17818

Abstract

This study aims to uncover parallels between the data issues observed in ORBIS and those in SABI, highlighting the need for cautious interpretation of both databases. By examining differences between entities present in SABI and those absent, insights into database representativeness are gained. Results indicate that SABI, like ORBIS, may not fully represent Spain's business population. Furthermore, analysis suggests that newer, smaller companies are less likely to appear in SABI, impacting data comprehensiveness. Extending this analysis, further variables will be explored to enhance understanding. This study underscores the importance of careful data scrutiny and the consideration of database limitations in research and decision-making processes.

Keywords: *SABI; BORME; ORBIS; bias; data; Python.*

1. Introduction

The Sistema de Análisis de Balances Ibéricos (SABI) database, developed by Bureau van Dijk, is a financial database and analysis tool that provides information on companies in Spain and Portugal. It is widely used by businesses, researchers, financial analysts, and professionals to access comprehensive information about companies. Research using SABI illustrates its broad applicability across many different topics and studies (Martínez-Matute & Urtasun, 2022; Rizov et al. 2022; Sánchez-Infante, et al. 2020).

Similar to SABI, other databases cover different geographic areas. Bureau van Dijk also offers FAME (UK and Ireland), AIDA (Italy), DAFNE (Germany) and ORBIS (Global), among others. They present a comprehensive reach and frequently serve as a proxy for the total firm population in research (Opazo-Basáez et al., 2024; Garcés-Galdeano et al., 2024; Martinez-Sanchez and Lahoz-Leo, 2018). However, these databases do not exhaustively represent the corporate landscape, as they offer limited coverage, especially for small and micro firms (Bajgar

et al. 2020, Almunia et al. 2018, Pinto Ribeiro et al. 2010). Therefore, the practice of considering companies listed there as the population may overlook the fact that they constitute a sample rather than a complete census. This distinction is crucial for accurately interpreting findings derived from its data, highlighting the need for awareness regarding its scope and limitations in research.

Simultaneously, the BORME[1] is the official gazette for business registrations and updates in Spain, providing a legal record of new companies, modifications, and terminations. As a primary source of official business information, BORME plays a crucial role in maintaining transparency and up-to-date records of the business landscape in Spain. Although the information that BORME contains for each firm is limited, the intersection of records between SABI and BORME reveals a unique opportunity to assess and enhance the completeness of business databases. The challenge lies in BORME's format—a text-based PDF without tabular data—which complicates direct comparisons with SABI's structured database. Overcoming this barrier requires innovative data extraction and analysis methods, emphasizing the importance of advanced technological solutions in bridging the information gap between these two essential resources.

Our study aims to identify and describe companies that appear in the BORME but are not found in the SABI database. While the representativeness of databases like ORBIS, Bloomberg SPLC, and Compustat has been examined in prior research (Liu 2020; Pinto Ribeiro et al. 2010; Bajgar et al. 2020; Culot et al. 2023), studies specifically focused on SABI are lacking. We investigate the differences between BORME and SABI to reveal key characteristics of omitted companies, such as their year of establishment and year of dissolution. Our goal is to understand how these characteristics affect the completeness and reliability of the SABI database. By identifying potential biases or omissions, our results provide valuable insights for improving the quality of economic analyses, policymaking, and business strategy development that rely on such databases.

The remainder of the paper is structured as follows. Section 2 reviews some literature about the data quality issues in business databases. Section 3 explains the methods followed to obtain the data and their description. Section 4 the results obtained from the data with a brief explanation. Finally, Section 5 presents some concluding remarks.

2. Related work

The challenges related to data quality in business databases, particularly those like Orbis, are critical considerations for researchers relying on such sources for comprehensive firm-level

[1] Boletín Oficial del Registro Mercantil

information. This analysis delves into the prevalent issues, ranging from missing values to data errors and biases, affecting the reliability and representativeness of datasets like Orbis.

Biases in databases are widely discussed. Survivorship bias and selection bias were reported in Datastream and Orbis (Andrikopoulos et al., 2007; Ince & Porter, 2006; Kalemli-Ozcan et al., 2019). Biases can happen for various reasons. Overstatement by a statistical measure or index can result in an upward bias; when observations are excluded from the sample due to a selection rule other than random sampling, it can create selection bias and survivorship bias is an example of the selection bias driven by the disproportionate exclusion of stocks that were delisted over time.

Missing values are one of the most prevalent data quality problems. In Orbis, missing value can also occur due to the cap on the amount of data allowed to be downloaded (Kalemli-Ozcan et al., 2019) where the research emphasizes that "In spite of the extensive use of the Orbis database for research, firm-level data downloaded from this database are not nationally representative…" and finally they provide a guide for researchers on how to download and organize the data such that it ends up being nationally representative or comes close to being so. Also researchers sometimes take special procedures or filters to exclude missing values from the research sample. However, this practice may inevitably create omission bias or selection biases (Elton et al., 2001; Weiß & Muhlnickel, 2014; Liu, 2020). Dropping all observations that contain missing values is a naïve strategy and can have a marked effect on the statistical power of the tests (Hribar, 2016, p. 63) and excluding these missing values can create misleading results. This great number of missing values may make a database not usable for specific research (Francis et al., 2016; Lee, 2017).

There are also other problems like data errors (Monasterolo et al., 2017), inconsistencies (Kalemli-Ozcan et al., 2019), static header data issue or vintage issue (Kalemli-Ozcan et al., 2019) and reporting time issues (Kalemli-Ozcan et al., 2019). Also, it is worth noting that when making comparisons and using ORBIS data, caution is required, especially when dealing with different countries. This is because some variables or data may not refer exactly to the same thing, as mentioned on the ORBIS website: "Our reports are in standardized formats to accommodate regional variations in filing regulations and accountancy practices..." Although this is secondary to the issues with the data itself. Therefore, when conducting studies with these types of databases, careful attention must be paid to all the potential issues they may present. It is crucial to approach research with a clear understanding of the challenges inherent in these databases. This issue is not exclusive to specific databases; similar problems can also be encountered in SABI.

3. Data

3.1. Data sources

Data for companies established between 2010 and 2023 have been collected from two sources: SABI and BORME. SABI is a commercial database and provides data in a convenient tabular format. After selecting those established in Spain within the period under study, a list of 1,911,775 companies was retrieved.

However, BORME is a set of daily publications in PDF format, available on the official website. By means of web crawling and scraping techniques, those publications were downloaded and converted to text. The structure of each publication is a list of entries in the registry where each entry corresponds to a company. Each entry is associated with specific registry events, such as the establishment of the company, a capital increase or a declaration of bankruptcy.

To construct the dataset, around 100,000 publications were downloaded and 9,956,791 registry entries corresponding to 3,051,505 companies were processed. After filtering out companies not established in the selected period, 2,917,784 entries associated with 1,298,056 companies were kept. Each registry entry was transformed into tabular format and grouped by company. Data from SABI and BORME were finally merged to create the dataset used in the analysis.

3.2. Data description

The final dataset consisted of 1,298,056 companies. Table 1 describes the variables used in this paper, although some others representing various registry events were also collected and processed.

The downloaded database from SABI that is ultimately used consists of a total of 1,911,775 companies and the parameters/variables selected include the company name, the NIF code of the company, the BvD number that identifies the company in SABI, the province, and finally, the date of incorporation. After downloading and loading the database, certain process where applied to adjust to what is needed just like selecting only companies between the years 2010 and 2023, text elimination, date transformation, text transformations, etc.

As for the BORME database, it consists of much larger dataframes, just like mentioned before in Barcelona it has 1,5 million rows and a 33 columns. Each row in the dataframe corresponds to a record of an event that has occurred and was registered by the company in a certain province on a certain date. It's worth noting that a company can appear multiple times in the dataframe. And some columns are “nombre actual”, “constitución”, “fecha”, “provincia”, “extinción”, etc. Which some of them are dates data types or numbers but most of them are texts. This dataframe undergoes transformations, where, as mentioned earlier, the records are grouped by companies (using the current names of the companies). The variables have been transformed from texts to

numerics and it been grouped by other parameters, such as the registration year or the year of incorporation.

After the merging of both datasets, we end up with a dataset where each row corresponds to a company, and each row contains different variables. For this time only“ extinción” and others variables, such as the year of incorporation for each company or whether that company is present or not in the SABI database are used.

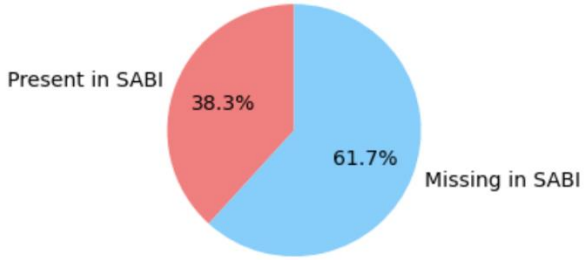
Table 1. The variables used for the study explained and the type of the variable.

Variable	Definition	Units
Year	The year when the event was registered.	Date (Year)
Year of incorporation	The year when the company was created/ incorporated.	Date (Year)
Sabi	Whether the company is present in SABI or not.	Binary
Company	Name of the company	Text
Dissolution	If the company in that year has or not dissolved.	Binary

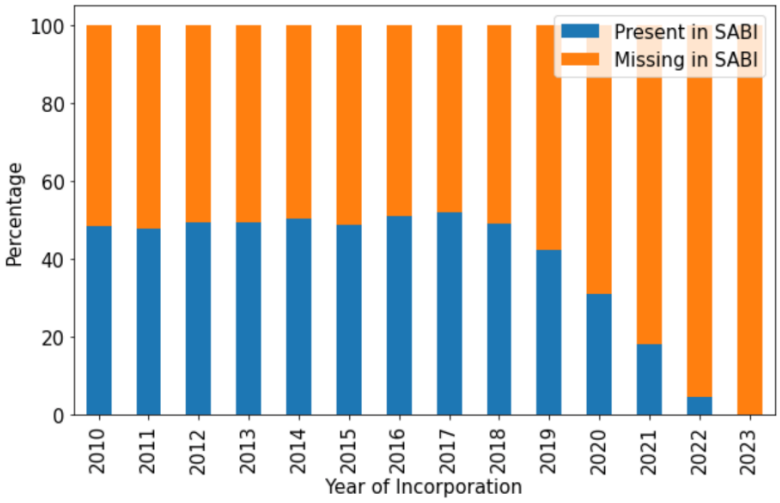
4. Results

4.1. First insight

In the initial study, a broad and generic analysis was conducted. To achieve this, a pie chart was created to observe the proportion of companies that are present in SABI and those that are not. Then, conducting a more in-depth study, a stacked bar chart where the X-axis reflects the year of incorporation of the companies, and the Y-axis represents the percentage. The 100% on the Y-axis corresponds to all the companies incorporated during that year. Each stacked bar is divided into two parts: the blue section denotes those present in SABI, and the orange section represents those absent in SABI.



a)



b)

Figure 1. The proportion of companies present in SABI and missing in SABI (a) in general, and (b) by year of incorporation.

As the graph depicts, in red, approximately 40% represents the percentage of companies registered in the SABI database, while in blue, around 60%, represents those not found in SABI. And observing the second graph, a distinction can be made between two periods. The first period spans from 2010 to 2018, where the proportion remains relatively constant, close to 50%. However, in the second period starting from 2018, there is a decreasing trend each year, with an average decrease of 7.2% per year, approaching the present.

4.2. Dissolution

Having explored the companies in a general context, we will delve deeper by studying more specific variables. In this instance, the focus is on analyzing the dissolution variable, aiming to identify potential patterns or differences between companies present and absent in SABI. To address this analysis, we have created two complementary graphs. The first is a temporal graph

representing in blue the total dissolution of SABI-listed companies and in orange those not in SABI. The second graph is a stacked bar chart illustrating the proportion of both categories. It is essential to note that the years in these graphs do not represent the year of each company's establishment, but rather the year in which the dissolution event is recorded.

It can be observed from the first graph that, in both cases, the growth is positive. This is not surprising since, as mentioned earlier regarding the years, as the year increases, so does the number of companies, which can explain this growth. Although both show an increase, companies not present in SABI tend to dissolve more in the early years, excluding 2010 (a special case). Analyzing both graphs, and more prominently in the second one, it can be noted that as the year is more recent, SABI-listed companies also tend to dissolve more, approaching the percentage of those not in SABI. This phenomenon occurs with an average growth of around 1.41% per year. Although there is a balance in the most recent years, in subsequent years, companies outside SABI tend to dissolve more.

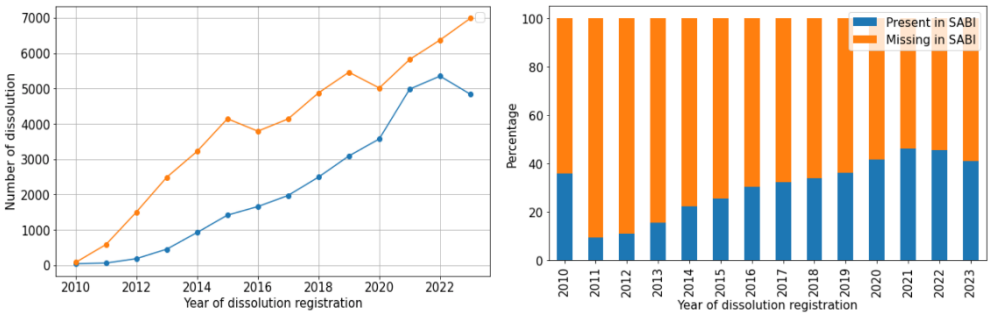


Figure 2. Quantity and proportion of dissolutions per year for companies that are present in SABI and those that are missing from SABI.

5. Conclusion

From the obtained results, it can be concluded that the issue present in ORBIS is also evident in SABI. The companies collected in SABI cannot be considered the complete population of Spain, emphasizing the need for caution when analyzing this data. Also observing the second graph from the figure 1, the decline may be attributed to the fact that, as the year of incorporation becomes more recent, companies are generally smaller and newer, making them less likely to be included in SABI. That can also be corroborated after studying the extinction where not in SABI companies tend to extinct since for smaller and newer companies is more challenging to sustain. Although we cannot conclusively state this until we have explored more variables, which will be done subsequently and not included here due to the limited number of pages that can be included.

References

Andrikopoulos, P., Daynes, A., Pagas, P., & Latimer, D. (2007). UK market, financial databases and evidence of bias (Occasional Paper Series Paper No. 79). http://www.dmu.ac.uk/documents/business-and-law-documents/business/occasional-papers/paper79ukmarketfinancialdatabasesandrikopoulos.pdf

Annaert, J., Buelens, F., & Riva, A. (2016). Financial history databases: Old data, old issues, new insights? In D. Chambers & E. Dimson (Eds.). Financial market history (pp. 44–65). Charlottesville, VA: CFA Institute Research Foundation.

Arbelo, A., Arbelo-Pérez, M. & Pérez-Gómez, P. (2022). Are SMEs less efficient? A Bayesian approach to addressing heterogeneity across firms. Small Bus Econ 58, 1915–1929. https://doi.org/10.1007/s11187-021-00489-2

Bajgar, M., et al. (2020), Coverage and representativeness of Orbis data. OECD Science, Technology and Industry Working Papers, No. 2020/06. https://doi.org/10.1787/c7bdaa03-en

Blanco-Mazagatos, V., Romero-Merino, M.E. & Santamaría-Mariscal, M. et al. One more piece of the family firm debt puzzle: the influence of socioemotional wealth dimensions. Small Bus Econ (2024). https://doi.org/10.1007/s11187-024-00881-8

Bostwick, E. D., Lambert, S. L., & Donelan, J. G. (2016). A wrench in the COGS: An analysis of the differences between cost of goods sold as reported in Compustat and in the financial statements. Accounting Horizons, 30(2), 177–193. https://doi.org/10.2308/acch-51336

Chychyla, R., & Kogan, A. (2014). Does Compustat data standardization improve bankruptcy prediction models? Social Science Research Network. http://ssrn.com/abstract=2406136

Elton, E. J., Gruber, M. J., & Blake, C. R. (2001). A first look at the accuracy of the CRSP Mutual Fund Database and a comparison of the CRSP and Morningstar Mutual Fund Databases. The Journal of Finance, 56(6), 2415–2430. https://doi.org/10.1111/0022-1082.00410

Francis, R. N., Mubako, G., & Olsen, L. (2016). Archival research considerations for CRSP data. Social Science Research Network. https://ssrn.com/abstract=2608273

Hribar, P. (2016). Do Compustat financial statement data articulate? Journal of Financial Reporting, 1(1), 61–63. https://doi.org/10.2308/jfir-51329

Ince, O. S., & Porter, R. B. (2006). Individual equity return data from Thomson Datastream: Handle with care! Journal of Financial Research, 29(4), 463–479. https://doi.org/10.1111/j.1475-6803.2006.00189.x

Kalemli-Ozcan, S., Sorensen, B., Villegas-Sanchez, C., Volosovych, V., & Yesiltas, S. (2019). How to construct nationally representative firm level data from the Orbis Global Database: New facts and aggregate implications (No. w21558). National Bureau of Economic Research. https://www.nber.org/papers/w21558.pdf

Lee, J. (2017). How do firms choose their debt types? http://www.fmaconferences.org/Boston/P1_201608.pdf

Liu, G. (2020). Data quality problems troubling business and financial researchers: A literature review and synthetic analysis. Journal of Business & Finance Librarianship, 25(3-4), 315-371. https://doi.org/10.1080/08963568.2020.1847555

Martín-Rojas, R., Garrido-Moreno, A. & García-Morales, V. J. (2020). Fostering Corporate Entrepreneurship with the use of social media tools. Journal of Business Research, 112, 396-412. https://doi.org/10.1016/j.jbusres.2019.11.072

Martínez-Matute, M., & Urtasun, A. (2022). Uncertainty and firms' labour decisions. Evidence from European countries. Applied Economics, 25(1), 220-241. https://doi.org/10.1080/15140326.2021.2007724

Monasterolo, I., Battiston, S., Janetos, A. C., & Zheng, Z. (2017). Vulnerable yet relevant: The two dimensions of climate-related financial disclosure. Climatic Change, 145(3–4), 495–507. https://doi.org/10.1007/s10584-017-2095-9

Opazo-Basáez, M., Monroy-Osorio, J. C. & Marić, J. (2024). Evaluating the effect of green technological innovations on organizational and environmental performance: A treble innovation approach. Technovation, 129, 102885. https://doi.org/10.1016/j.technovation.2023.102885

Pinto Ribeiro, S., Menghinello, S., & De Backer, K. (2010). The OECD ORBIS Database: Responding to the Need for Firm-Level Micro-Data in the OECD. OECD Statistics Working Papers, No. 2010/01. https://doi.org/10.1787/5kmhds8mzj8w-en

Rico, M., Pandit, N.R. & Puig, F. (2021). SME insolvency, bankruptcy, and survival: an examination of retrenchment strategies. Small Bus Econ 57, 111–126. https://doi.org/10.1007/s11187-019-00293-z

Rizov, M., Vecchi, M. & Domenech, J. (2022). Going online: Forecasting the impact of websites on productivity and market structure. Technological Forecasting and Social Change, 184, 121959. https://doi.org/10.1016/j.techfore.2022.121959

Sánchez-Infante Hernández, J. P., Yañez-Araque, B. & Moreno-García, J. (2020). Moderating effect of firm size on the influence of corporate social responsibility in the economic performance of micro-, small- and medium-sized enterprises. Technological Forecasting and Social Change, 151, 119774. https://doi.org/10.1016/j.techfore.2019.119774

Sánchez-Vidal, F. J., Hernández-Robles, M. & Mínguez-Vera, A. (2020): Financial conservatism fosters job creation during economic crises. Applied Economics 52:45, 4913-4926 https://doi.org/10.1080/00036846.2020.1751053

Segarra, A., Callejón, M. (2002). New Firms' Survival and Market Turbulence: New Evidence from Spain. Review of Industrial Organization, 20, 1–14. https://doi.org/10.1023/A:1013309928700

Weiß, G. N., & Mühlnickel, J. (2014). Why do some insurers become systemically relevant?. Journal of Financial Stability, 13, 95–117. https://doi.org/10.1016/j.jfs.2014.05.001

Yáñez-Araque, B., Sánchez-Infante Hernández, J. B., Gutiérrez-Broncano, S. & Jiménez-Estévez, P. (2021). Corporate social responsibility in micro-, small- and medium-sized enterprises: Multigroup analysis of family vs. nonfamily firms. Journal of Business Research, 124, 581-592. https://doi.org/10.1016/j.jbusres.2020.10.023

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17432

Electoral abstention and information sources among undergraduate university students

Jorge Mora Rojo[1,2], **Jose Manuel Tomás**[1], **Víctor Yeste**[3,4], **Eduardo Cebrián**[2,5]

[1]Department of Methodology for the Behavioral Sciences, Universitat de València, Spain, [2]School of Social Sciences, Universidad Europea de Valencia, Spain, [3]Department of Applied Statistics and Operational Research, and Quality, Universitat Politècnica de València, Spain, [4]School of Science, Engineering and Design (STEAM), Universidad Europea de Valencia, Spain, [5]Department of Economics and Social Sciences, Universitat Politècnica de València, Spain

How to cite: Mora, J.; Tomás, J. M.; Yeste, V.; Cebrián, E. 2024. Electoral abstention and information sources among undergraduate university students. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17432

Abstract

A quantitative research study was conducted to examine electoral abstention among undergraduate students in Valencia during Spain's General Elections held in July 2023. Data were collected through a survey based on a questionnaire designed specifically for this purpose, focusing on the electoral behavior, socioeconomic and demographic profiles, and information sources. A multivariate statistic model was used to explain the likelihood of students non-voting. Key findings indicate that abstention it is not related to a demographic profile. However, a negative correlation with abstention was found among information sources such as party websites/social media, YouTube, printed press and TV, whereas a positive relationship was found with the use of blogs and forums.

Keywords: *digital citizenship; social media; electoral abstention; online platforms; electoral programs; logistic regression.*

1. Introduction

Our study on electoral behavior focuses on abstention in the last General Elections in Spain, held in July 2023, with university students in the city of Valencia as our target population. Electoral participation represents an important indicator of the health of a democracy, of social commitment (Varela et al., 2015), and of the level of voters' information in a society.

During this period, individuals form their comprehension of society and develop the capacity to understand and engage in politics, acquiring this knowledge through diverse socialization agents, with a notable emphasis on media channels. The initial formation of political attitudes definines one's political identity, which exhibits enduring traits that vary in intensity over the

years (Varela et al., 2015). Furthermore, young electoral population engage in elections less than adults (Espí Hernández, 2020).

The work of Shah et al. (2009) highlights the importance of information in socialization and the formation of civic citizens. They identify the four primary agents of socialization—family communication, debate activities in educational institutions, mass media, and peer discussion—as the foundation of communicative competence, which are interconnected. Thus, the media play a crucial role in providing information and understanding to young citizens about public affairs and electoral behavior (Goh et al., 2011; Moeller et al., 2014; Wlezien & Soroka, 2019).

We can distinguish between traditional media such as TV, newspapers, and radio, and media that have emerged with the penetration of digital technology including news websites, political party websites, social media, blogs, forums and podcasts (Soe, 2018).

The variety of sources can be a strong indicator of the interest in political life. If someone seeks information from different sources, they are likely to engage more deeply in understanding the various aspects of political events (Carlson, 2019). Consulted authors have observed a growing trend towards the use of emerging platforms, especially social media for consuming information, while there is simultaneou decrease in the use of traditional and digital media for this purpose (Kaid et al., 2007; Moeller et al., 2014). Similarly, Catalina-García et al. (2018) note that university students, categorized as the so-called true digital natives identified or Generation Z (Childers & Boatwright, 2021), preferentially use the Internet for information. However, there is a significant tendency for nearly all to combine on line news consumption with traditional media.

The utilization of online media allows the development of debating skills. Discussing political events reported in the news serves as a more accurate predictor of political involvement than mere exposure to the news (Shah et al., 2009). Furthermore, websites promote political participation by enabling interactive connections between citizens and parties (Norris, 2003).

2. Objectives

This study aims to explain the likelihood of electoral abstention among university students in the city of Valencia during the General Elections of July 2023. It focuses on the impact of information sources while considering socioeconomic and demographic characteristics as control variables.

Specific objectives are: to investigate wether online information channels are better predictors of electoral participation than offline channels, and to examine the socioeconomic and demographic profile of the undergraduate students.

3. Methodology

Regarding electoral behavior, our interest lies in explaining participation in the last Parliamentary and Senate elections in July of 2023 in Spain. We consider a binary indicator: abstention and participation. Explanatory variables are defined in Table 1.

Table 1. Exogen explanatory variables. Source: Own elaboration.

Control variables: sex, year of birth, household income			
Dimension	**Indicators**	**Item**	**Values**
Information	*Platforms and social media:* sources use to follow current social news and political events	I1.What are the communication channels you use to get information?	Multiple choice: printed press, digital press, radio, TV, ChatGPT, SSNN, conversations
		I2. Do you use any of the following online media to get information?	Multiple choice: Facebook, Instagram, Youtube, Twitter, Tiktok, blogs, forums.
	Electoral campaign: reading of the political party programs	E2. Have you read the program of any party?	Closed response: Yes/No.
		E3. Indicate whether you have consulted the website and/or SSNN of any party	Closed response: Both/only website/ only social media/ None.

Sex is a binary variable; Birth year has been recoded into age in years; Household income is measured as an ordinal variable taking values: ≤900€, 901-1.200€, 1.201-1.800€, 1.801-2.400€, 2.401-3.000€, 3.001-4.500€, 4.501-6.000€, > 6.000€; E3 was recoded into a binary variable: indicating wether individuals have consulted the website and/or social media of any political party and individuals or have not. Categories from items I1 and I2 were considered for the generalized linear regression as dummy variables indicating presence or absence. E2 and E3 were also considered as indicator of presence or absence.

Our research employed a non-experimental, cross-sectional design with a non-probabilistic intentional sample consisting of 598 undergraduate students across three universities in Valencia. Data were collected through a survey built ad hoc, with Google Forms platform. The survey was launched on September 7th, 2023 and concluded on the 25th of November 2023. The Universidad Europea de Valencia (UEV) ethics committee approved the research. The questionnaire was validated by 14 expert judges. The original questionnaire consisted of 55 questions, a final selection of 35 question was made. Inclusion criteria were: participants being enrolled in a university undergraduate program and to give informed consent in accordance with the Helsinki Declaration. Participants received detailed explanation about their involvement and

the research objectives. At the UEV, educators directly distributed the survey to students, while at the Universitat de València (UV) and the Universitat Politècnica de València (UPV), students were approached on their way out of the different campus premises.The size sample for the universities surveyed can be found in Figure 1.

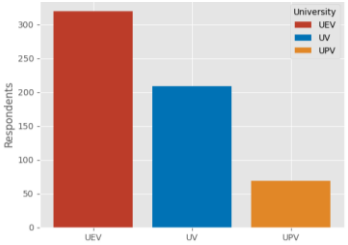


Figure 1. Distribution of survey respondents by university. Source: Own

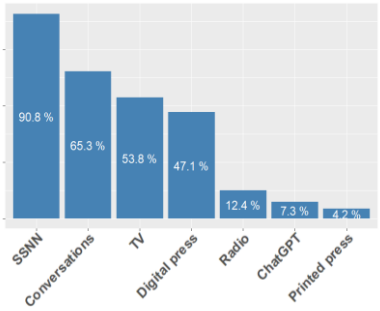


Figure 2. Distribution information chanels. Source: Own elaboration.

A binomial logistic regression (BLR) was utilized to analyze the relationship between abstention and the explanatory variables detailed in Table 1. Two models have been run, the first includes variables I1, E2 and E3, while in the second model the SSNN category from I1 is substituted by the different online sources chosen by the student as specified in E2. We did not apply any variable selection method, since the aim of this research is to explore which variables present a statistically significant association with abstentionism. A p-value lower than .05 has been considered statistically significant in this study.

3. Results

Firstly, it was noted that 13.5% of the sample, 64 out of 472 individuals, reported not having voted in the last General Spanish Elections.

The initial hypothesis is that there is a significant relationship between the likelihood of abstention and the varoius media channels and sources through which students consult and gather information.

Model 1, detailed in Table 2, explores abstention using dummy variables for the different types of media, such as Digital Press, Radio or TV, along with control variables including Sex, Birth Year and Household Income of the respondent. Model's likelihood-ratio (LLR) test p-value, is .001<.01, Indicating with with more than 99% of confidence that the model fits the data better than the null model, which includes the intercept as the only parameter. The model achieves ae pseudo-R2 of 0.1. It was found that only TV and Political party websites hold a significant negative relationship with abstention, which provides evidence that students who gather

information through either of these two channels are less likely to abstain. Moreover, social media and political party programs also present a weakly significant relationship, p-value <.1. This relationship is negative in the case of social media, but positive in the case of political party programs, which evidences that those students who use political party programs as a mean of information present a higher likelihood of abstention among their users. Among the control variables, none showed are significant association on the abstention of undergraduate students in Valencia.

Table 2. Logistic regression Model 1 with media consulted by the student. Source: Own elaboration.

Model 1				**Model 2**			
Variable	**Coef**	**OR**	**p**	**Variable**	**Coef**	**OR**	**p**
Sex	0.22	1,25	.49	Sex	0.50	1.65	.15
Birth year	-0.07	0,93	.37	Birth year	-0.05	0.95	.61
Household income	-0.07	0,93	.33	Household income	-0.07	0.93	.35
Printed press	-0.89	0,41	.43	Printed press	-0.68	0.51	.54
Digital press	-0.22	0,80	.51	Digital press	-0.35	0.70	.33
Radio	0.43	1,54	.37	Radio	0.27	1.31	.60
TV	-0.72	0,49	.03*	TV	-0.78	0.46	.03*
SSNN	-0.83	0,44	.08.	Facebook	0.67	1.95	.26
				Instagram	0.75	2.12	.08.
				YouTube	-0.85	0.43	.02*
				Twitter	-0.39	0.68	.25
				Tiktok	-0.48	0.62	.20
				Blog	0.80	2.23	.13
				Forums	1.02	2.77	.04*
ChatGPT	-0.14	0,87	.82	ChatGPT	-0.35	0.70	.59
Friends and family	-0.03	0,97	.93	Friends and family	-0.13	0.88	.70
Party programs	0.75	2,12	.07.	Party programs	0.63	1.88	.15
Party webs / SSNN	-0.93	0,39	.03*	Party webs / SSNN	-1.2	0.30	.01**
Pseudo R-Squared	0.107			Pseudo R-Squared	0.159		
LLR P-Value	.001**			LLR p-value	<.001***		

. p-value < .1 *p-value < .05 **p-value < .01

Because of the marginally significant relationship found for SSNN in Model 1, p-value = .08, Model 2 elaborates on this by specifying different SSNN channels as independent variables, as shown on Table 2. The thought behind this specification is that the weak significant relationship might be dued from the relevance of certain social media platforms over others.

The LLR p-value< .001 for Model 2 and the pseudo-R^2=.16, so it can be concluded with more than 99% confidence, Model 2 fits better the behaviour of abstencion than the null model and it performs better than Model 1.

Analysis reveals that once all the social media platforms are considered individually, YouTube, Forums and Instagram show significant or weakly significant effects, while the rest of social media platform do not present significant effects on the abstention of undergraduate students in Valencia. Specifically, YouTube presents a negative relationship, which indicates that those students gathering information through YouTube are less likely to abstain. Conversely, student utilizing forums or Instagram as a mean for political information present a higher likelihood of abstention. Additionally control variables remain not significant in Model 2.

4. Conclusions and Discussion

The low abstention rate, 13.5%, suggests that the university population is more engaged commitment in electoral participation compared to older cohorts. If it is state that younger age ranges show higher electoral abstention, then from our results we can infer that they also show lower abstention compared to non-university youth.

We have chosen not to explore the potential causal link between information source usage and electoral participation, as such relationships could be spurious. Although, only 4 out of the 18 independent variables examined in model 2, were found to be significant, it is also important to reflect the strong effect achieved in certains predictors. Specifically, regarding the relationship between social media and turnout: for each person who did not go to the polls and and received information via Youtube, approximately 2.3 chose to vote. Similarly, voter participation was a 61% and 47% higher among TikTok and X users. Conversaly, we highlight the positive relationship obtanined in the social networks from the Meta company, Facebook and Instagram, finding a ratio around double the individuals who chose not to vote; the use of blogs and forums also reached ratios higher than 2, nearly the triple among forum users.

Regarding traditional media, the strongest relationship with participation is observed in the the activity of consulting parties' website or social networks, which showed a ratio of 3.3 voters for every one who abstained; Additionally, TV has also a positive relationship with electoral participation, classifying 2.2 people who participated for every one who did not.

Regardless of the effectiveness of communication programs, political parties are betting on electoral campaigns through online channels. The literature review suggest that young people have displaced traditional channels by online media as a source of information: Based on our findings we recommend that political parties prioritize engagement through YouTube, TikTok and X as social media, and also focus on enhancing their websites and partisan social networks.

Furthermore, young individuals who read the electoral programs also approach double the people who abstained. The use of blogs, forums, and the consultation of electoral programs is a more active attitude of information search and comprehension compared to the rest of the channels in which individuals are passive spectators of information consumption. In this way, students who have not voted have a more critical behavior and attitude. These results agree with the hypothesis posed by Boulliane, who argues that a greater frequency of use of social networks, blogs, or forums does not increase participation in voting.

On our final analysis on socioeconomic and demographic factors revealed that none of these variables were found to explain statistically participation behavior, corroborating Moeller's findings that age and gender do not have significant effect. Althouh Age was analyzed as a quantitative variable with its linear and cuadratic effect, but al so as a dummy variable, a null effect was always found. The introduction of gender as a control variable revealed that men were 65% more likely to abstain from voting. This finding supports Catalina-García et al.'s research, which noted differences in information consumption patterns between genders.

References

Carlson, T.N. (2019). Through the Grapevine: Informational Consequences of Interpersonal Political Communication. *American Political Science Review, 113*(2), 325-339. https://doi.org/10.1017/S000305541900008X

Catalina-García, B., López De Ayala, M. C., & Martín, R. (2018). Medios sociales y la participación política y cívica de los jóvenes. Una revisión del debate en torno a la ciudadanía digital. *Doxa Comunicación. Revista interdisciplinar de estudios de comunicación y ciencias sociales, 27*, 81-97. https://doi.org/10.31921/doxacom.n27a4

Childers, C., & Boatwright, B. (2021). Do digital natives recognize digital influence? Generational differences and understanding of social media influencers. *Journal of Current Issues & Research in Advertising, 42*(4), 425-442.

Espí, A. (2019). Protagonistas del cambio: identidades políticas y participación electoral de los jóvenes en España, 1982-2016. *Acciones e Investigaciones Sociales*, (40), 193-217. https://doi.org/10.26754/ojs_ais/ais.2019404202

Goh, K.Y., Hui, K.L., & Png, I. P. (2011). Newspaper reports and consumer choice: Evidence from the do not call registry. *Management Science*, *57*(9), 1640-1654. https://doi.org/10.1287/mnsc.1110.1392

Kaid, L.L., McKinney, M.S., & Tedesco, J.C. (2007). Introduction: Political Information Efficacy and Young Voters. *American Behavioral Scientist*, *50*(9), 1093-1111. https://doi.org/10.1177/0002764207300040

Moeller, J., De Vreese, C., Esser, F., & Kunz, R. (2014). Pathway to political participation: The influence of online and offline news media on internal efficacy and turnout of first-time voters. *American Behavioral Scientist*, *58*(5), 689-700. https://doi.org/10.1177/0002764213515220

Norris, P. (2003). Preaching to the converted? Pluralism, participation and party websites. *Party politics*, *9*(1), 21-45. https://doi.org/10.1177/135406880391003

Pallarés, F., Riba, C., & Fraile, M. (2007). Variables socioestructurales y comportamiento electoral en las elecciones generales españolas. Una perspectiva evolutiva 1979-2000. *Revista de Estudios Políticos*, (135), 109-158.

Shah, D.V., McLeod, J.M., & Lee, N. (2009). Communication competence as a foundation for civic competence: Processes of socialization into citizenship. *Political Communication*, *26*(1), 102-117. https://doi.org/10.1080/10584600802710384

Soe, Y. (2018). Understanding politics more thoroughly: How highly engaged young citizens use the Internet for civic knowledge integration. *First Monday*, *23*(6). https://doi.org/10.5210/fm.v23i6.7923

Varela, E., Martínez, M. L., & Cumsille, P. (2015). ¿Es la participación política convencional un indicador del compromiso cívico de los jóvenes? *Universitas Psychologica*, *14*(2), 731. https://doi.org/10.11144/Javeriana.upsy14-2.eppc

Wlezien, C., & Soroka, S. (2019). Mass media and electoral preferences during the 2016 US presidential race. *Political Behavior*, *41*(4), 945-970. https://doi.org/10.1007/s11109-018-9478-0

Zazueta, I.M.S., & Cortez, W.W. (2014). *Determinantes de la participación electoral en México*. *Estudios sociológicos*, *32*(95), 323-353. https://www.jstor.org/stable/24368103

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17441

Read between the headlines: Can news data predict inflation?

Alan Chester Arcin, Ma. Ellysah Joy Guliman, Genna Paola Centeno, Jacqueline Margaux Herbo, Sanjeev Parmanand, Cherrie Mapa
Department of Economic Research, Bangko Sentral ng Pilipinas, Philippines.

How to cite: Arcin, A.; Guliman M.; Centeno G.; Herbo J.; Parmanand S.; Mapa C. 2024. Read between the headlines: Can news data predict inflation? In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17441

Abstract

Big data and machine learning applications are increasingly gaining traction in central bank operations. Among others, central banks have tapped big data in their nowcasting exercises. In this study, we construct inflation news indices and examine if such indices can help predict inflation. The indices are developed using lexicon-based sentiment analysis refined using reinforcement learning and supervised machine learning methods, particularly artificial neural networks and long short-term memory models. These indices are then used as additional feature variables in time series models and machine learning models for nowcasting regional and nationwide inflation in the Philippines. We find empirical evidence that our constructed inflation news indices can improve the predictive capability of these forecasting models.

Keywords: *Inflation; news; nowcasting; sentiment analysis; machine learning*

1. Introduction

Big data and machine learning (ML) based applications are increasingly gaining traction in central bank operations. The 2020 survey conducted by the Irving Fisher Committee on Central Bank Statistics of the Bank for International Settlements indicates that 80 percent of the central banks surveyed utilize big data for their operations, up from just 30 percent in 2015. Among various types of applications, the survey showed that central banks have used big data in their nowcasting exercises. This study adds to this existing body of empirical studies by tapping novel data sources for nowcasting regional and nationwide inflation in the Philippines. Inflation nowcasting and forecasting is important for central banks like the Bangko Sentral ng Pilipinas (BSP) given that it operates an inflation targeting framework for monetary policy formulation.

While Beck et al (2023) and Macias and Stelmasiak (2019) used online price data to nowcast inflation, our research examines if textual data such as online news articles could be useful for inflation nowcasting. This study also builds upon an earlier work by Gabriel et al (2020) for

regional inflation nowcasting in the Philippines which employed machine learning techniques but did not incorporate big data. We opt to explore online news data as an indicator for inflation trends as major developments that are relevant for inflation are likely captured by news reports.

To answer our research question, we employ a two-step approach. First, we build inflation news indices which provide information on whether there are more reports of increasing/higher inflation vis-à-vis decreasing/lower inflation in the news media. We take a comprehensive approach and explore two methods of text analytics: (a) lexicon-based sentiment analysis refined by reinforcement learning and (b) ML-based approach using neural networks and long short-term memory models. Second, we employ these indices as additional feature variables in time series models and in ML models for inflation nowcasting.

Overall, we find empirical evidence that our indices are useful for nowcasting regional and nationwide inflation in the Philippines. The news-based indices have information content that help improve the forecasting accuracy of models that utilize the said indices as additional feature variables relative to models that did not.

2. Constructing the Inflation News Indices

Our proposed inflation news indices (INIs) harness the power of online news media to capture real-time information on price trends of goods and services and even other factors that could affect overall inflation. This section outlines the steps in the construction of our own lexicon- and ML-based INIs.

2.1. Data sources and annotation

News articles for this study are sourced from media outlets in the Philippines that allowed us to web scrape the data from their websites' business, finance, or economy sections.[1] Our news database contains articles from different time periods but the common period when we have data from all media outlets is from January 2018 to present. We further filtered our data to news articles containing any of the words: "inflation", "price" and "prices". A total of 3,000 randomly selected articles are annotated. These are manually labeled as 1, -1, or 0, indicating increase, decrease or no change in inflation, respectively. The annotation of sample sentences is a crucial first step to generate and to implement further enhancements of the lexicon- and ML-based INIs.

[1] These media outlets include Business Mirror, Business World, Manila Bulletin, Manila Standard and Philippine Daily Inquirer.

2.2. Lexicon with reinforcement learning method

This method involves the creation of a set of words and usage of predefined set of rules. In this study, the words are grouped to indicate *increase* or *decrease* in inflation. The negation rule is used but instead tweaked to indicate *no change*.[2]

Following Church and Hanks (1990), pointwise mutual information (PMI) was leveraged to identify which words are classified as *increase*/*decrease*. The overall score $Score_{PMI}$ for a given word *w* is the difference between PMI score of the word with respect to *increase* and *decrease*. Words having $Score_{PMI} > 0$ and $Score_{PMI} < 0$ are therefore assigned under *increase* and *decrease* wordlists, respectively.

The resulting wordlists is manually evaluated for suitability and reinforcement learning is used to further refine the wordlists. In particular, we will use Q-learning (a type of reinforcement learning (Watkins and Dayan, 1992)) to retrieve the optimal set of words under *increase* and *decrease* that maximizes the Macro-F1 score of the test set within a reasonable amount of time.

2.3. ML-based method using neural networks and long short-term memory models

This method relies on supervised ML models trained using our manually annotated sentences, The sentences are transformed using Word2Vec and are fed into artificial neural networks (ANNs) such as multilayer perceptron (MLP) and Bidirectional Long Short-Term Memory (BiLSTM) models. We test several model configurations and explore the use of Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al, 2002) to create synthetic entries and remedy data scarcity concerns. For the purposes of this study, all ANNs are designed to classify three output classes: *increase*, *decrease*, and *no change*. We then compare the Macro-F1 scores of the different ANNs.

2.4. Index construction and evaluation

To construct the INIs, we evaluate the lexicon and trained models based on their accuracy and Macro-F1 scores. Table 1 presents a summary of evaluation of different lexicons and ML models. The results suggest that INIs derived using lexicon refined by reinforcement learning (hereafter referred to as INI-RL) and using BiLSTM (32) (hereafter referred to as INI-BiLSTM) have the best accuracy and Macro-F1 scores are therefore preferred methods for index construction. These are then used to score the sentences in the news articles with scores aggregated per article and indexed for the month. Afterwards, the INIs are normalized to a mean of 100. Thus, any estimate above 100 suggests that relative to average, news about increasing

[2] Negation rule applies when a word is preceded by a negation word (e.g., “not” or “never”)

inflation outnumber reports of declining inflation while any estimate below 100 suggests otherwise.

Table 1. INI methods: Evaluation metrics for test set. Note: Number in parenthesis for ANNs refers to number of hidden units. Source: Authors' estimates

	Accuracy (In percent)	Macro-F1 score (In percent)
Initial lexicon	64.3	41.8
Lexicon with reinforcement learning	89.3	69.3
MLP	68.9	45.7
BiLSTM (16)	78.4	56.2
BiLSTM (32)	79.8	57.9

We find that the INI-RL and INI-BiLSTM tend to move together with correlation coefficient of 0.93. This provides assurance in our measurement of inflation data contained in news as two different methodologies tend to yield similar results.

3. Forecasting inflation using news data

The key research question in our research is to determine the usefulness of news data in nowcasting inflation in the Philippines. Following Gabriel et al (2020), we first estimate time series (TS) model and ML models without the INIs. We compare the forecast accuracy of these models based on their mean absolute error (MAE). We then augment these models with the INIs and perform another round of forecast evaluation to determine if the inclusion of INIs leads to improved forecasting capability.

In this study, we employ a TS and ML models for regression tasks, namely: Support Vector Machines (SVM), Gradient Boosted Machines (GBM) and Extremely Randomized Trees (EXT).

SVM (Cortes and Vapnik, 1995) finds a hyperplane that separates various classes while maximizing the distance between these classes. It uses kernel functions to map data into higher dimensions to expedite separation of classes. Well-known for solving complex problems, SVM performs well in small datasets but scales poorly.

GBM (Friedman, 2001) is an algorithm known for its speed and accuracy, with large datasets in particular. Popular in machine learning competitions, it is an ensemble model, initially starting and combining multiple weak decision trees. Sequentially, subsequent models improve on the previous ones and are evaluated on the loss function of the ensemble.

EXT (Geurts et al, 2006), also known as extra-trees, are similar to random forests (RF) where it fits a number of random decision trees. The main difference is ERT uses the entire dataset to generate trees while RF selects from different variations of data via bagging. This reduces variance and improves computational speed relative to RF.

To match the sample period for our INIs, we restrict all estimations from January 2018 to December 2023 with a train-test split of 80:20. This corresponds to January 2018 to September 2022 and October 2022 to December 2023 for the train and test sets, respectively. For all these models, the target variable is the year-on-year inflation rate.[3] We build individual models for each of the 16 regions and a separate model for nationwide inflation using the different modeling techniques. For our baseline models, we use the following as feature variables: autoregressive component (lag order of 2), previous month's month-on-month inflation and percentage change in year-on-year inflation rate per month to retrieve information embedded in the historical time series of inflation itself. Following Gabriel et al (2020), a shock variable *S* defined as the difference of the year-on-year inflation rate between the current month and the previous month is introduced as an additional variable in the baseline models. This is defined in Equation 1 below:

$$S_t = \begin{cases} 1, if\ YoY\pi_t > YoY\pi_{t-1} \\ 0, \quad otherwise \end{cases} \tag{1}$$

For the models with INIs, we add another shock variable S^{INI} defined in Equation 2:

$$S_t^{INI} = \begin{cases} 1, \ if\ INI_t > INI_{t-1} \\ -1, if\ INI_t < INI_{t-1} \\ 0, \quad otherwise \end{cases} \tag{2}$$

where INI_t refers to either INI-RL or INI-BiLSTM as a shock variable at month *t*.

4. Analysis of Results

This section presents the results of our inflation nowcasting exercises. As we are concerned mainly with out of sample forecasting performance, we only present the MAE from the test set. Table 2 displays the results for the baseline models. Similar to the findings of Gabriel et al (2020), we find evidence that the SVM outperforms the TS as it registers lower MAE across all regions. However, the GBM and EXT fail in this aspect. Given the weaker forecasting performance of GBT and EXT, we opt to no longer include these models in the assessment of models that will be augmented with the INIs.

[3] Data on regional and nationwide inflation for the Philippines are sourced from the Philippines Statistics Authority.

Table 2. Forecast Evaluation: MAE for test set (October 2022 – December 2023). Source: Authors' estimates.

	TS	SVM	GBT	EXT
NCR	0.49	0.47	0.85	0.82
CAR	0.62	0.40	0.55	0.62
R1	0.62	0.35	0.88	0.61
R2	0.60	0.46	0.61	0.68
R3	0.66	0.40	1.83	1.14
R4A	0.57	0.41	0.74	0.89
R4B	0.57	0.40	0.78	0.75
R5	0.64	0.43	0.50	0.45
R6	0.78	0.59	1.35	1.25
R7	0.44	0.48	0.47	0.52
R8	0.58	0.37	0.87	0.63
R9	0.98	0.72	0.76	0.55
R10	0.69	0.37	0.62	0.61
R11	0.59	0.40	0.69	0.68
R12	0.55	0.34	0.51	0.34
BARMM	0.68	0.37	0.90	0.44
R13	0.48	0.36	0.44	0.38
Philippines	0.44	0.34	0.85	0.65
Average	0.61	0.43	0.79	0.67

Table 3 shows the forecast evaluation of the TS and SVM models that are augmented with INI-RL and INI-BiLSTM. The results show that the models augmented with new-based indices have registered lower forecasting errors compared to baseline models. This holds true for either the INI-RL or the INI-BiLSTM. This provides empirical evidence that the INIs contain information that can help nowcast regional and nationwide inflation in the Philippines.

Table 3. Forecast evaluation: MAE for test set (October 2022 – September 2023). Source: Authors' estimates.

	Base SVM	SVM+INI-RL	SVM+INI-BiLSTM	TS	TS+INI-RL	TS+INI-BiLSTM
NCR	0.31	0.23	0.31	0.25	0.25	0.27
CAR	0.29	0.03	0.22	0.26	0.26	0.24
R1	0.35	0.31	0.13	0.31	0.26	0.28
R2	0.32	0.28	0.28	0.33	0.30	0.29
R3	1.24	0.13	0.10	0.34	0.32	0.31
R4A	0.33	0.23	0.26	0.34	0.27	0.28
R4B	0.40	0.07	0.45	0.34	0.27	0.25
R5	0.32	0.20	0.23	0.30	0.27	0.30
R6	0.44	0.37	0.36	0.35	0.36	0.34
R7	0.41	0.38	0.32	0.38	0.41	0.38
R8	0.34	0.29	0.31	0.27	0.24	0.24
R9	0.56	0.47	0.52	0.53	0.47	0.47
R10	0.22	0.25	0.21	0.24	0.23	0.23
R11	0.30	0.30	0.29	0.36	0.34	0.33
R12	0.35	0.29	0.36	0.36	0.38	0.38
BARMM	0.43	0.20	0.16	0.34	0.21	0.22
R13	0.19	0.34	0.22	0.20	0.24	0.22
Philippines	0.34	0.22	0.05	0.25	0.25	0.24
Average	0.40	0.26	0.27	0.32	0.29	0.29

5. Conclusion and Future Directions

Can news data help predict inflation? The short answer to this research question is yes. In this study, we provide empirical evidence that the inclusion of INIs can help improve the predictive capability of time series and SVM models for regional and nationwide inflation nowcasting in the Philippines. Our news-based indices have information content that are relevant for monitoring inflation trends. The models that we have built in this study can serve as complementary models for the suite of forecasting models maintained by the BSP.

Our findings contribute to the existing literature on the relevance of big data for nowcasting macroeconomic variables. Future plans on this exercise include the addition of other ML models for generating regional forecasts and combining multiple ML models as a ensemble of its own as another model for prediction.

References

Beck, G. W., Carstensen, K., Menz, J. O., Schnorrenberger, R., & Wieland, E. (2023). Nowcasting consumer price inflation using high-frequency scanner data: Evidence from Germany (No. 34/2023). Deutsche Bundesbank Discussion Paper.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. Computational linguistics, 16(1), 22-29.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20, 273-297.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Gabriel M., Bautista D., & Mapa C. (2020). Forecasting regional inflation in the Philippines using machine learning techniques: A new approach. Bangko Sentral ng Pilipinas Working Paper Series No. 2020-10.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine learning, 63, 3-42.

Macias, P., & Stelmasiak, D. (2019). Food inflation nowcasting with web scraped data (p. 302). Warsaw: Narodowy Bank Polski, Education & Publishing Department.

Mahadevaswamy, U. B., & Swathi, P. (2023). Sentiment analysis using bidirectional LSTM network. Procedia Computer Science, 218, 45-56.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26.

Serena, J. M., Tissot, B., Doerr, S., & Gambacorta, L. (2021). Use of big data sources and applications at central banks (No. 13). Bank for International Settlements.

Watkins, C. J., & Dayan, P. (1992). Q-learning. Machine learning, 8, 279-292.

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17503

Nowcasting food insecurity interest Google Trends data

Nicola Caravaggio[1], **Bia Carneiro**[2], **Giuliano Resce**[1]

[1]Department of Economics, University of Molise, Campobasso, Italy, [2]Bioversity International, Rome, Italy.

How to cite: Caravaggio, N.; Carneiro, B. ; Resce, G. 2024. Nowcasting food insecurity interest Google Trends data. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17503

Abstract

This research explores the potential of Google Trends (GT) data as a tool for generating a daily index of food insecurity at the national level, focusing on regions monitored by the Famine Early Warning Systems Network (FEWS NET) and the Global Fragility Act (GFA). Drawing inspiration from previous studies on GT's predictive capabilities, the authors employ Natural Language Processing (NLP) to analyse food security reporting from FEWS NET documents. We identify key predictors of food insecurity using a LASSO regression approach and construct a daily economic sentiment index (DESI) for each country. Unlike traditional methods, the study considers multiple languages and weights search terms based on LASSO coefficients. The resulting Synthetic Search Interest (SSI) index for food insecurity demonstrates a statistically significant correlation with FAO's share of the population in severe food insecurity, affirming GT's potential as a monitoring tool. The research contributes a novel methodology and insights into leveraging real-time data for early warnings in food security.

Keywords: *food insecurity, Google trends, early warnings, Natural Language Processing*

1. Introduction

Official Statistics indicators regarding food, nutrition, and livelihood security outcomes are typically only available with a reporting lag of several weeks and are often revised a few months later. Nowadays, several sources of data on real-time economic activity are available from private sector companies such as Alphabet. An example is Google Trends (GT), a real-time daily, weekly, and monthly index of the volume of user queries on Alphabet's search engine Google. Such user search patterns are often correlated with various socio-economic indicators and may be helpful for short-term prediction and nowcasting.

This research aims to provide a strong tool able to generate a daily index of food insecurity at national level by relying on GT data. Considering that threats to food security affect mostly

fragile and food insecure regions, the presented analysis has been developed on countries monitored by the Famine Early Warning Systems Network (FEWS NET) and the Global Fragility Act (GFA). Henceforth, GT may present an important tool for early warnings of food insecurity in highly stressed areas. Ginsberg et al. (2009) pioneered the use of GT in research studies. Their groundbreaking work showcased how GT could effectively monitor and forecast the progression of influenza ahead of the official reports from the Centers for Disease Control and Prevention (CDC) in the United States. Other authors have stressed the predictive potentiality of GT, especially regarding the present: "[w]e are not claiming that Google Trends data help predict the future. Rather we are claiming that Google Trends may help in predicting the present" (Choi & Varian, 2009) [p. 2].

2. The potentialities in the use of GT

Researchers are increasingly relying on GT under manifold circumstances (Jun, Yoo, & Choi, 2018). Choi and Varian (2009) have shown how GT data can help predict initial claims for unemployment benefits in the United States. Furthermore, Askitas and Zimmermann (2009) and Suhoy (2009) performed similar analyses stressing the potential of GT in Germany and Israel, respectively. Conversely, Nagao, Takeda, and Tanaka (2019) stressed some limitations in using GT to nowcast unemployment in the US. By working on retail, automotive, and home sales topics, Choi and Varian (2012) also demonstrated how seasonal autoregressive (AR) models and fixed-effects models that includes relevant GT variables tend to outperform models that exclude these predictors. Several studies used GT data within the financial market to predict, for example, the direction of stock market through neural networks trained with GT data (Fan, Chen, & Liao, 2021; Hu, Tang, Zhang, & Wang, 2018) or the identification of "early warnings signs" in financial markets (Petropoulos, Siakoulis, Stavroulakis, Lazaris, & Vlachogiannakis, 2022; Preis, Moat, & Stanley, 2013). Extensive application of GT data is also found within the field of epidemiology, for example, as source of real-time influenza surveillance (Broniatowski, Paul, & Dredze, 2013). Moreover, GT data is not confined solely to predictive areas but also as a support for user geolocation of Twitter (now X) data (Zola, Ragno, & Cortez, 2020). Nonetheless, these approaches and the general use of GT are not free from critiques (Cook, Conrad, Fowlkes, & Mohebbi, 2011; Lazer, Kennedy, King, & Vespignani, 2014).

During the outbreak of the Covid 19 pandemic, several studies investigated the phenomenon through the lens of GT (Kornellia & Syakurah, 2023). Kurian et al. (2020) evidenced the high correlation between Covid cases among US states and 10 keywords searched on GT; Liu et al. (2022) through a prophet model showed how GT of Covid related terms represented important predictors in investigating the number of cases among US states; Lampos et al. (2021) showed how online searches precede the number of confirmed cases and deaths by 16.7 and 22.1 days, respectively; Brunori and Resce (2020) relied on GT data to estimate a prediction model for the Italian case. Those studies generally demonstrate how online search data can be used to develop

complementary public health surveillance methods in conjunction (not substitution) with more established approaches.

3. Data and methodology

Our work starts from a Natural Language Processing (NLP) analysis of food security reporting based on 1,414 publicly available documents from FEWS NET, covering 33 countries. The development of a custom taxonomy of food insecurity identified three groups of topics, namely (i) hazard/shocks, (ii) food security indicators, and (iii) food, nutrition, and livelihood security outcomes, each in turn composed of different sub-topics. A topic-matrix was created for the occurrence of each sub-topic among the list of country-year reports. For the purposes of this study, we rely only on the first topic-group (hazards/shocks) as it considers terms which effectively represent predictors of food insecurity. For example, it makes a more lot more sense in case of a food crisis that people would search on Google for topics which may effectively be identified as shocks (*e.g.*, conflict, exchange rate) rather than more general, outcome-related terms associated to this phenomenon (*e.g.*, food production, hunger). Topics within the hazards/shocks theme have been divided into five groups: climate, conflict, markets, diseases, and governance. We considered a total number of 47 countries, mostly located in Sub-Saharan Africa over a 10-year period spanning from January 1, 2013, to December 31, 2023.

In the first step of our analysis, we identified the most important predictors of food insecurity through a data driven approach using a LASSO regression on the classification results from the NLP model. Then, we selected the top ten positive features after excluding those disease-related, such as Covid-19 or Ebola, due to their time or geographic specificity. The necessity to identify only the most important features is twofold: (i) through the LASSO we were able to determine a feature's importance with relative coefficients that we used eventually to weight our GT search terms; (ii) the necessity to automate the download procedure from GT lead to a necessary balancing between data granularity and feature selection[1].

After determining the core predictors of food insecurity represented in FEWS NET reporting, we followed the daily economic sentiment index (DESI) approach proposed by Eichenauer, Indergand, Martinez, and Sax (2022). For each country, a long-run frequency-consistent daily trend of food insecurity was constructed. However, our approach is different in two major aspects. First, rather than considering only one specific language, we preferred to retrieve word search data by considering all national languages of our sample of countries, in addition to the most widely spoken languages in the world not included among the national languages of the countries analysed. This means that in each country, the search for a specific word has been replicated for 25 different languages. Results were then aggregated in order to constitute, for

[1] The download procedure has been carried out with R using the package gtrendsR.

each topic, a country-specific index. Second, to aggregate search terms into a single indicator, we preferred to weight each topic by the corresponding scaled coefficients obtained from the LASSO and then average results, rather than use a principal component analysis (PCA). In fact, we noticed that trying to create a synthetic index starting from trends of numerous words (up to 10), the PCA may not represent the most suitable tool since it often aggregates into an index of disinterest, hence assuming a pattern generally opposite to what is shown by their components.

4. Final outcome

The final time series constructed for each country represents a synthetic search interest (SSI) index for food insecurity based on GT data. In the final step of our procedure, we applied some further weights allowing for cross-country comparison. The first weight has been constructed by considering the worldwide GT interest of each topic while the second one relied on yearly data of internet penetration (WB, 2024) (inverse) to compensate for the digital divide among countries.

Lastly, we validate our data by performing a comparison between our SSI index and the share of population in severe food insecurity provided by FAO (2024). Results show a statistically significant and positive correlation between the two data suggesting GT may effectively help in representing a tool for monitoring food insecurity trends worldwide. The whole procedure adopted is graphically synthesized in *Figure **1*** while an example of SSI for Tanzania is shown in *Figure **2***.

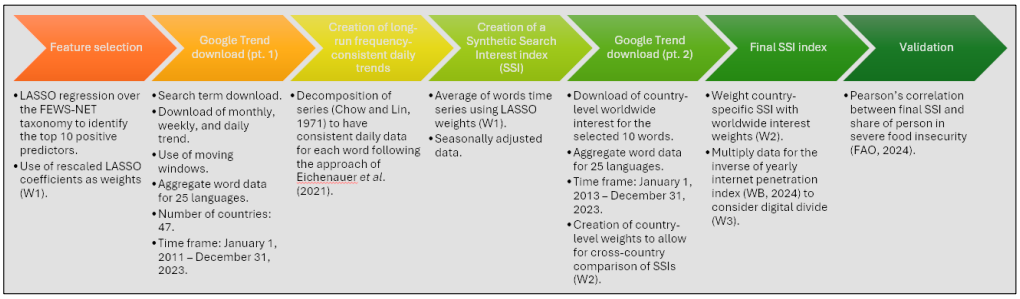


Figure 1. Cross-country synthetic search interest (SSI) creation procedure.

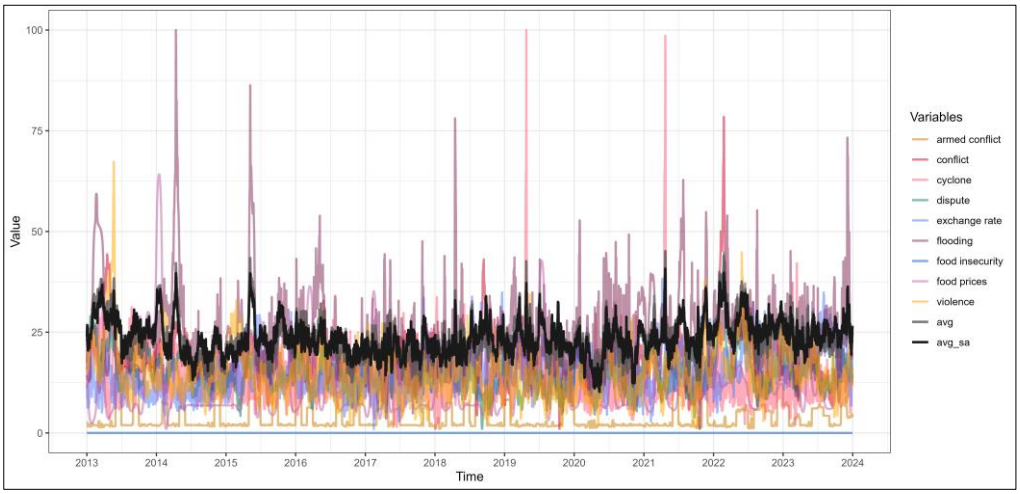


Figure 2. Example of synthetic search interest (SSI) for Tanzania.

References

Askitas, N., & Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. Applied Economics Quarterly, 55(2), 107–120.

Broniatowski, D. A., Paul, M. J., & Dredze, M. (2013). National and local influenza surveillance through twitter: an analysis of the 2012-2013 influenza epidemic. PloS one, 8(12), e83672.

Brunori, P., & Resce, G. (2020). Searching for the peak google trends and the covid-19 outbreak in italy. (SERIES working papers, No. 04/2020)

Choi, H., & Varian, H. (2009). Predicting initial claims for unemployment benefits. Google Inc, 1 (2009), 1–5.

Choi, H., & Varian, H. (2012). Predicting the present with google trends. Economic record, 88, 2–9.

Cook, S., Conrad, C., Fowlkes, A. L., & Mohebbi, M. H. (2011). Assessing google flu trends performance in the United States during the 2009 influenza virus a (h1n1) pandemic. PloS one, 6(8), e23610.

Eichenauer, V. Z., Indergand, R., Mart´ınez, I. Z., & Sax, C. (2022). Obtaining consistent time series from google trends. Economic Inquiry, 60(2), 694–705.

Fan, M.-H., Chen, M.-Y., & Liao, E.-C. (2021). A deep learning approach for financial market prediction: Utilization of google trends and keywords. Granular Computing, 6, 207–216.

FAO. (2024). FAOSTAT. Rome, IT. (Food and Agriculture Organization of the United Nations.

Retrieved from: http://www.fao.org/faostat/en/#data [Accessed February 1, 2024])

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. Nature, 457(7232), 1012–1014.

Hu, H., Tang, L., Zhang, S., & Wang, H. (2018). Predicting the direction of stock markets using optimized neural networks with google trends. Neurocomputing, 285, 188–195.

Jun, S.-P., Yoo, H. S., & Choi, S. (2018). Ten years of research change using google trends: From the perspective of big data utilizations and applications. Technological forecasting and social change, 130, 69–87.

Kornellia, E., & Syakurah, R. A. (2023). Use of google trends database during the covid-19 pandemic: systematic review. Multidisciplinary Reviews, 6(2), 2023017–2023017.

Kurian, S. J., Alvi, M. A., Ting, H. H., Storlie, C., Wilson, P. M., Shah, N. D., . . . others (2020).

Correlations between covid-19 cases and google trends data in the United States: A state-by-state analysis. In Mayo clinic proceedings (Vol. 95, pp. 2370–2381).

Lampos, V., Majumder, M. S., Yom-Tov, E., Edelstein, M., Moura, S., Hamada, Y., . . . Cox, I. J. (2021). Tracking covid-19 using online search. NPJ digital medicine, 4(1), 17.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of google flu: traps in big data analysis. Science, 343(6176), 1203–1205.

Liu, Z., Jiang, Z., Kip, G., Snigdha, K., Xu, J., Wu, X., . . . Schultz, T. (2022). An infodemiological framework for tracking the spread of sars-cov-2 using integrated public data. Pattern Recognition Letters, 158, 133–140.

Nagao, S., Takeda, F., & Tanaka, R. (2019). Nowcasting of the us unemployment rate using google trends. Finance Research Letters, 30, 103–109.

Petropoulos, A., Siakoulis, V., Stavroulakis, E., Lazaris, P., & Vlachogiannakis, N. (2022). Employing google trends and deep learning in forecasting financial market turbulence. Journal of Behavioral Finance, 23(3), 353–365.

Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using google trends. Scientific reports, 3(1), 1684.

Suhoy, T. (2009). Query indices and a 2008 downturn: Israeli data (Tech. Rep.). Bank of Israel.

WB. (2024). World Development Indicators. Washington, D.C., US. (World Bank. Retrieved from: https://databank.worldbank.org/source/world-development-indicators [Accessed February 1, 2024])

Zola, P., Ragno, C., & Cortez, P. (2020). A google trends spatial clustering approach for a worldwide twitter user geolocation. Information Processing & Management, 57(6), 102312.

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17817

Mapping Circular Economy in Spain with LinkedIn data

Theodoros Daglis[1,3], George Tsironis[2], Pavlos Fafalios[1], Konstantinos P. Tsagkarakis[1]

[1] School of Production Engineering and Management, Technical University of Crete, Greece, [2]Department of Environmental Engineering, Democritus University of Thrace, Greece, [3] School of Applied Economics and Social Sciences, Department of Agricultural Economics and Rural Development, Agricultural University of Athens, Greece.

How to cite: Daglis, T.; Tsironis G.; Fafalios P.; Tsagarakis K.P. 2024. Mapping Circular Economy in Spain with LinkedIn data. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17817

Abstract

This paper presents a quantitative approach that captures the LinkedIn information that can be derived regarding companies and jobs that operate in a circular economy context. To do so, we perform this analysis for the regions of Spain, providing metrics for the standardization of the companies identified, for measurement reasons, while job posts and other information is analyzed. Finally, heatmaps and other graphs demonstrate the distribution of LinkedIn information for the various regions of Spain, concluding that Spain's regions do not perform in a homogenous way in LinkedIn regarding the circular economy.

Keywords: *Circular economy; networking; social media; online job posts; LinkedIn.*

1. Introduction

Social media provide alternative and useful information regarding individuals, job posts, companies, sentiment, and other features, capable of benefiting enthusiasts and stakeholders (Utz, 2016). LinkedIn is a professional platform that operates globally, therefore a plethora of information can be derived regarding the business activity (Davis et al., 2020).

Transforming from linear to circular business models within a company is vital for enhancing business performance since it promotes resource efficiency by reducing waste and optimizing the use of materials, leading to cost savings and improved operational efficiency while boosting innovation in product design, manufacturing processes, and business models, which can drive revenue growth and market differentiation. Consequently, examining the circular economy (CE) through LinkedIn is crucial as it facilitates professional networking, highlighting business opportunities, and enabling knowledge sharing.

LinkedIn is a social network that focuses on business and employment aspects. In this regard, information can be derived for several analyses, including content effects of sales revenue on business-to-business services (Mora Cortez et al., 2023), individuals' participation impact on social capital formation in the context of LinkedIn (Mashayekhi and Head, 2022), or the relationship between the number of followers on the funds raised by companies (Banerji and Reimer, 2019).

LinkedIn has been already used for the investigation of several aspects of companies that operate in various frameworks. For example, Daglis and Tsagarakis (2024) utilize LinkedIn data to examine U.S. healthcare companies operating within a COVID-19 framework, revealing distinctive characteristics and activities primarily in the "Health, wellness, and fitness" sector. Noteworthy, there are also a few studies already examining CE aspects through LinkedIn information. Tsironis et al. (2022) utilized LinkedIn data to assess the engagement of companies in circular economy activities in the EU, while Daglis et al. (2023) used LinkedIn data to measure circular economy interest and examine the relationship between keywords related to the circular economy and sustainability, demonstrating statistical significance in how circular economy keywords affect sustainability ones. Moreover, Tsironis and Tsagarakis (2023) discuss circular economy principles in the apparel, fashion, and textiles industry sectors based on data from companies' LinkedIn profiles, highlighting the emergence of companies worldwide, with a focus on recycling, repairing, reusing, and other related activities. Finally, Tsironis et al. (2024) leveraged LinkedIn data to explore global CE business activities, revealing insights into the geographical distribution of these companies, industry sectors, employee counts, followers, and foundation years. This study highlights the significant increase in new CE companies over the last decade and identifies prevailing strategies like reuse, reduce, and recycle, evident in companies' profiles and interactions.

Despite this subject's significance, not all mentioned techniques have been collectively applied to a region, therefore, we investigate the performance of Spanish companies in LinkedIn within the circular economy framework. Acknowledging this gap in the literature, this paper provides an empirical analysis that describes the Spanish performance in a country and regional context. The results indicate that the Spanish regions do not operate homogenously in a CE context.

2. Data & Methodology

To analyze corporate engagement in Circular Economy (CE) across different regions of Spain, we start by identifying the number of companies involved in CE activities in each region. Next, we determine the total number of companies operating in each region. By dividing the number of CE-engaged companies by the total number of companies in each region, we derive a metric that allows for direct comparison of CE engagement across regions. The formula used is as follows:

$$\frac{circular\ economy{-}related\ companies}{total\ number\ of\ companies} \quad (1)$$

Similar indicators can be derived by checking on the number of personal profiles or jobs in the regions. Furthermore, we searched for LinkedIn jobs using the query term "circular economy" and the location 'Spain', and then applied a set of filters provided by LinkedIn for a better understanding of the circular economy job market in Spain. We ran the query on 09/03/2024 at 22:50. We decided to use the English term "circular economy" instead of the Spanish term "economía circular" because the former returns much more results. The LinkedIn search service returned 125 jobs. Out of these jobs, 49 (39.2%) were posted the past week and 101 (80.8%) the past month. As regards the language of the job description, the large majority of the jobs are described in Spanish (84%) and the remaining in English (16%). Finally, we have searched for the industries, the year of establishment, and the specialties, entered by the selected companies.

3. Emperical Findings

We first present the companies heatmap of the regions of Spain. The distribution comes first for Catalonia, followed by the Community of Madrid, then Basque country, then comes the Balearic Islands, followed by the Valencian Community, then Aragon, Region de Murcia, Galicia, Castilla and Leon, Principality of Asturias, Castilla-La Mancha, Andalusia, and finally the Canary Islands. Based on the heatmap, the northeastern Spanish regions indicate more CE-related companies, while the south and west, have the least.

Regarding the experience level of the job postings, we notice that the majority of jobs concern mid-senior level (24.8%), followed by entry-level (20%) and associate (18.4%). Internship-level jobs occupy 10.4% of all jobs, while only 2.4% (3 jobs) are of the more senior 'director' level and 1.6% (2 jobs) of 'executive' level.

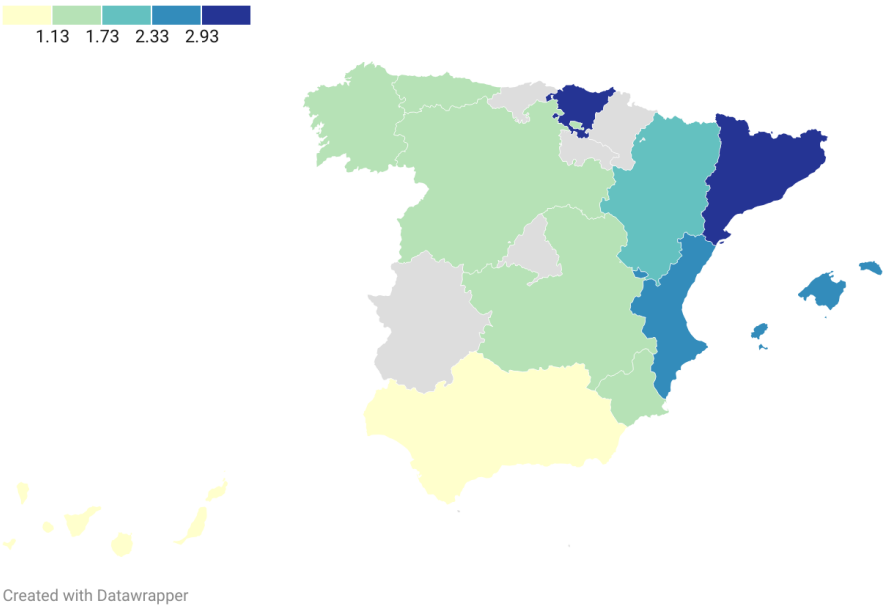


Figure 1. Heatmap for the CE companies' engagement for the Spanish regions.

Regarding the mode of operation, jobs offered are 54.4% on-site, 44.8% hybrid, and the remaining 0.8% (just one job) remote. Regarding the job type distribution, the large majority of jobs concern full-time employment (83.2%), followed by internship (7.2%) and contract (6.4%). Part-time employment is offered by only 2 jobs (1.6%). The most frequent locations of the job postings are Barcelona (19 jobs / 15.2%) and Madrid (14 jobs / 11.2%), followed by San Sebastian (9 jobs / 7.2%), Manresa (5 jobs / 4%) and Tarragona (4 jobs / 3.2%).

Figure 2 (left panel) shows the most frequent industries of the jobs. As expected, the majority of jobs are in 'Environmental Services' (13.6%), followed by 'Research Services' (10.4%), 'Engineering Services' (6.4%), 'Business Consulting and Services' (6.4%), and 'Technology, Information and Internet' (6.4%). It is also interesting that there are 7 jobs (5.6%) in the 'Oil and Gas' industry. Figure 2 (right panel) shows the most frequent job functions. We see that 'Engineering' is the most frequent function (26.4%), followed by 'Information Technology' (21.6%) and Management (16.8%). It is worth noting that oil and gas, construction, and industrial machinery manufacturing industries turn to circular economy, indicating that circular economy is desirable for various industries and fields, as well.

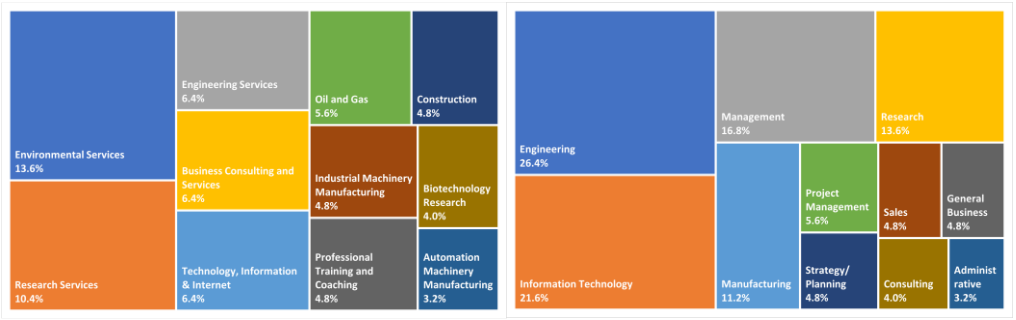


Figure 2. Job industries (left), and Job functions (right)

Based on Figure 3, the environmental services industry sector comes first, followed by renewables & environment, then comes information technology & services, plastics, higher education, packaging & containers, mechanical or industrial engineering, and many more. Note that LinkedIn company profiles are allowed to register only one Industry from a predefined list in English.

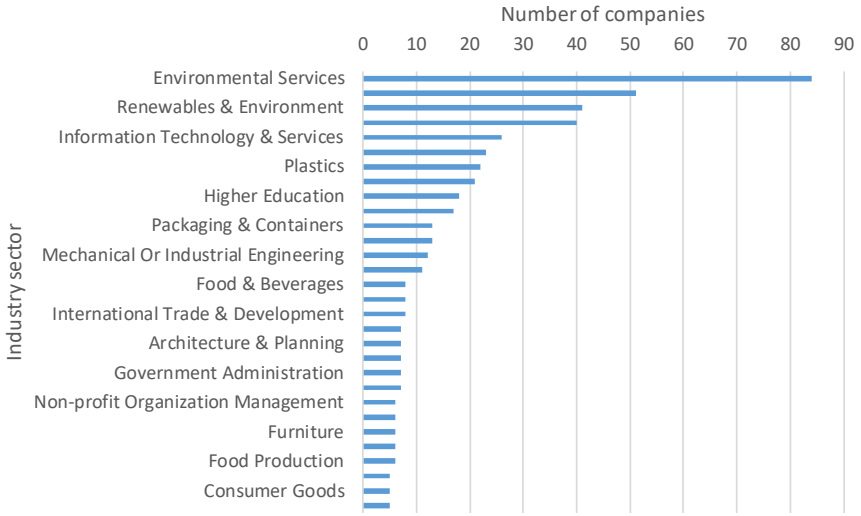


Figure 3. Industry sector frequency

The year with the highest number of companies' foundations is 2020, followed by 2021, and then 2018. After the year 2020 a decreasing trend is evidenced, more details about the historical evolution of company profiles related to CE are displayed in Figure 4.

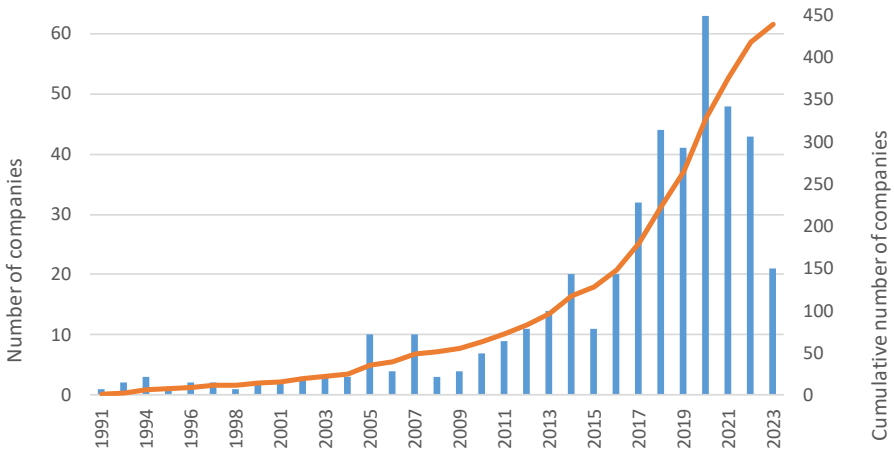


Figure 4. Year of establishment distribution.

Additionally, we investigated the descriptive statistics available for two significant metrics bound to the examined profiles, the followers and the number of LinkedIn registered employees. More analytically, followers indicate an average of 4037.5 and a standard error of 835.43, while staff count is 106.9, and 33.08, respectively (Table 1).

Table 1. Descriptive statistics of followers and staff

Statistics	Followers	Staff
Mean	4037.5	106.9
Standard Error	835.43	33.08
Median	425	3
Standard Deviation	20446.78	809.57
Kurtosis	110.54	239.21
Skewness	10.03	14.71
Minimum	0	0
Maximum	266494	14540
Sum	2418461	64012

Finally, we present the wordcloud for specialties (Figure 5) for the companies identified. Note that for each profile it is possible to enter from none to 20 specialties, which is equivalent to keywords. This is open to any number of words or languages. It is possible to get additional information from the descriptions of the companies, which is an open text. An initial assessment based on a word cloud is provided in Figure 6. These world clouds provide an initial assessment of the dominant activities, for example in the description section of the companies there is frequent mention to solution, product, service, innovation, material, and sustainability. There is however place for further cleaning and treatment to obtain a final classification.

Figure 5. Wordcloud for the specialties of the companies

Figure 6. Wordcloud for the description of the companies

4. Conclusions

Big data from social media can give essential information on topics of interest. Through this work, we provide an example of how data can be retrieved and analyzed for the circular economy, which can also be extended to any topic, focusing on any desired region or country, as well. Such an approach can provide important insights into the dynamics of circular economy attributes regarding various aspects of the domain of employment, at a regional, or global level. Our work provides important insights for stakeholders and decision-makers, who can benefit from trends and geographical indicators developed from this data. Finally, our work can be

expanded to a higher level of regional breakdown, scrutinizing the LinkedIn data per each region of Spain, and also providing a comparison, to demonstrate trends and probable future directions.

References

Banerji, D., & Reimer, T. (2019). Startup founders and their LinkedIn connections: Are well-connected entrepreneurs more successful? *Computers in Human Behavior*, 90, 46–52. https://doi.org/10.1016/j.chb.2018.08.033

Daglis, T., & Tsagarakis, K. P. (2024). A LinkedIn-based analysis of the U.S. dynamic adaptations in healthcare during the COVID-19 pandemic. Healthcare Analytics, 5, 100291. https://doi.org/10.1016/j.health.2023.100291

Daglis, T., Tsironis, G., & Tsagarakis, K. P. (2023). Data mining techniques for the investigation of the circular economy and sustainability relationship. Resources, Conservation & Recycling Advances, 19, 200151. https://doi.org/10.1016/j.rcradv.2023.200151

Davis, J., Wolff, H.-G., Forret, M. L., & Sullivan, S. E. (2020). Networking via LinkedIn: An examination of usage and career benefits. *Journal of Vocational Behavior*, 118, 103396. https://doi.org/10.1016/j.jvb.2020.103396

Mashayekhi, M., & Head, M. (2022). Developing social capital through professionally oriented social network sites. *Information & Management*, 59(6), 103664. https://doi.org/10.1016/j.im.2022.103664

Mora Cortez, R., Johnston, W. J., & Ghosh Dastidar, A. (2023). Managing the content of LinkedIn posts: Influence on B2B customer engagement and sales? *Journal of Business Research*, 155, 113388. https://doi.org/10.1016/j.jbusres.2022.113388

Tsironis, G., & Tsagarakis, K. P. (2023). Global online networking for circular economy companies in fashion, apparel, and textiles industries, the LinkedIn platform. Current Opinion in Green and Sustainable Chemistry, 41, 100809. https://doi.org/10.1016/j.cogsc.2023.100809

Tsironis, G., Daglis, T., & Tsagarakis, K. P. (2022). Social media and EU companies' engagement in circular economy: A LinkedIn approach. Sustainable Production and Consumption, 32, 802–816. https://doi.org/10.1016/j.spc.2022.06.006

Tsironis, G., Daglis, T., & Tsagarakis, K. P. (2024). The 21 most practiced RE-s of Circular Economy from LinkedIn Company Profiles on a Global Scale. Resources, Conservation & Recycling Advances, 200202. https://doi.org/10.1016/j.rcradv.2024.200202

Tsironis, G., Daglis, T., & Tsagarakis, K. P. (2024). The 21 most practiced RE-s of Circular Economy from LinkedIn Company Profiles on a Global Scale. Resources, Conservation & Recycling Advances, 200202. https://doi.org/10.1016/j.rcradv.2024.200202

Utz, S. (2016). Is LinkedIn making you more successful? The informational benefits derived from public social media. *New Media & Society*, 18(11), 2685–2702. https://doi.org/10.1177/1461444815604143

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17695

Enhancing Conflict Mediation Research: Introducing the Innovative Global Peace Actors Database (GLO-PAD)

Elisa D'Amico[1], **Mateja Peter**[2]

[1]School of International Relations, University of St Andrews, United Kingdom, [2]School of International Relations, University of St Andrews, United Kingdom

How to cite: D'Amico, E. & Peter, M. 2024. Enhancing Conflict Mediation Research: Introducing the Innovative Global Peace Actors Database (GLO-PAD). In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17695

Abstract

In contemporary conflict resolution research, integrating diverse data sources and innovative methodologies is crucial for understanding mediation events and actors. This paper presents the Global Peace Actors Database (GLO-PAD), an event-based mediation database built through a pioneering and layered data collection process. Leveraging various sources such as the Barcelona Peace Talks in Focus Report, International Crisis Group's Crisis Watch, PA-X Peace Agreements Database, and reports from international organizations, GLO-PAD offers a nuanced view of mediation dynamics. Unlike previous datasets, GLO-PAD employs a large-scale web scraping approach, integrating the universe of potential mediation event sources to capture granular mediation event data worldwide from 1988 to present. It transcends political negotiations to encompass all conflict-related mediation, thus broadening temporal and spatial scope of previous mediation data attempts. GLO-PAD assesses mediator bias, identifies emerging actors, and utilizes semi-automated techniques for comprehensive data collection. This innovative approach addresses mediation's increasing fragmentation, providing valuable insights for conflict resolution scholarship and practice.

Keywords: *Conflict Mediation, Peacebuilding, Data Collection, Global Peace Actors Database, Semi-Automation, Innovative Methodologies.*

1. Introduction

Effective conflict resolution and peacebuilding efforts require a nuanced understanding of mediation events and the diverse actors involved. The Global Peace Actors Database (GLO-

PAD)[1] represents a pioneering initiative to compile and analyze mediation data from various sources, offering insights into the complexities of peace processes. This paper outlines the methodology behind GLO-PAD and its implications for advancing conflict resolution research.

The scope of this study encompasses all mediation events related to armed conflicts worldwide, as documented by the Uppsala Conflict Data Program (UCDP), spanning from 1988 to the present day. This comprehensive temporal and spatial scope distinguishes our dataset from previous efforts in conflict mediation research.

2. Contributions

Unlike previous datasets such as the African Peace Processes (APP), which focused solely on Africa, or the UCDP Third Party Actors dataset, which concluded in the early 2000s, our dataset covers mediation events from across the globe over the past three decades. This broader spatial and temporal scope allows for a more comprehensive understanding of conflict mediation dynamics on a global scale.

Previous datasets often overlooked mediation events beyond political negotiations or peace agreements. Our dataset aims to capture a wider range of mediation events, including economic, humanitarian, and political interventions. By doing so, we provide a more nuanced understanding of the various approaches to conflict resolution. Conflict mediation is undergoing a fragmentation evolution, with an increasing number of new actors and actor types entering the mediation realm, as well as an increasing of mediation events in a broader peace process. Our project sheds light on these emerging actors and their diverse roles in conflict resolution. For example, Figure 1 illustrates how various actors engaged in the Sudan Peace Process collaborate and their levels of involvement. Understanding these dynamics is essential for adapting mediation strategies to evolving conflict and peace landscapes.

[1] Funded by the Foreign, Commonwealth and Development Office (FCDO) of the United Kingdom

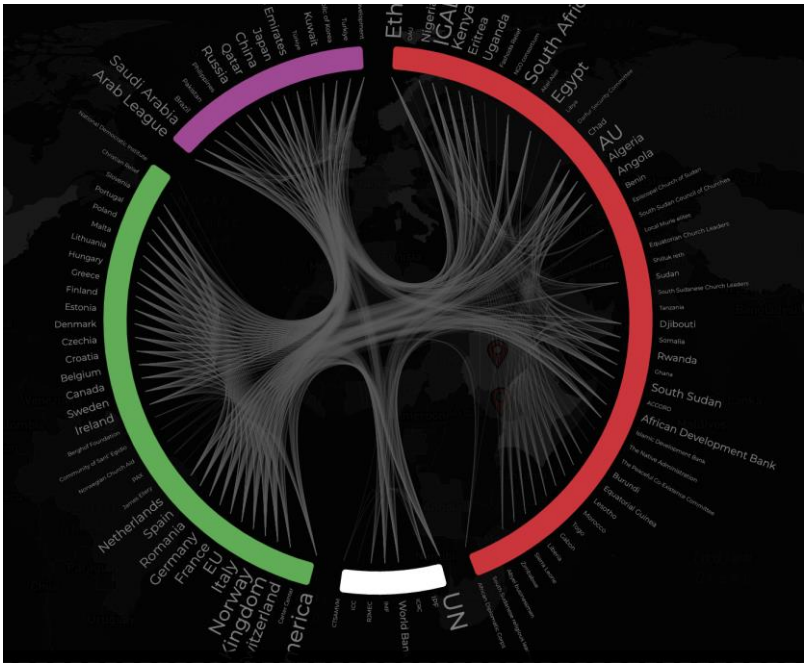


Figure 1. Third Party Actor Involvement Networks in the Sudan Conflict (1988-present). Data comes from the GLO-PAD visualization platform. The larger the actor name, the more mediation events they were engaged in. Red indicates regional actors; green indicates "global north" actors; white indicates international organizations; and purple indicates all "other" actors.

Another significant contribution of our study is the evaluation of mediator bias towards one of the conflict parties. Previous datasets failed to address this aspect, yet it plays a crucial role in shaping the outcomes of mediation efforts. By examining the level of bias exhibited by mediators, we enhance our understanding of the complexities involved in conflict resolution processes.

Our approach utilizes large-scale web crawling to capture extremely granular mediation event data from local news sources (among a vast array of other sources). This allows us to identify mediation attempts and actors that are not be captured by existing datasets or former data gathering methods. The inclusion of these previously overlooked actors is crucial for the development of conflict resolution scholarship and the formulation of effective peacebuilding strategies. For example, previous data fails to capture the levels of involvement in the conflict mediation process in Sudan from actors such as China, Russia, and smaller regional actors, whereas our data captures these comprehensively (see Figure 2).

Figure 2. Mapped Levels of State Actor Involvement in the Sudan Conflict (1988-present). Data comes from the GLO-PAD visualization platform. The "whiter" the country, the more mediation events they were engaged in. Pins identify points of conflict being referenced.

3. The Layered Approach to Data Collection

GLO-PAD employs a layered approach to data collection, beginning with the compilation of existing sources such as reports from Barcelona, the International Crisis Group, PA-X Peace Agreements Database, and international organizations. These sources provide a foundational framework for identifying key stakeholders and mediation events within conflicts. The following steps unfold in turn by gathering data on one conflict-country at a time in order to a careful and in-depth extraction of various actors and processes.

3.1. Step 1: Compilation of Existing Sources

The initial phase of data collection involves compiling a comprehensive list of existing sources that provide insights into conflict mediation efforts. This process begins with gathering information from institutions and organizations such as Universitat Autònoma de Barcelona and the International Crisis Group (ICG). These sources offer valuable reports and analyses that shed light on key actors, stakeholders, and mediation events within conflict zones.

The Peace Talks in Focus Report from Universitat Autònoma de Barcelona serves as a foundational resource by offering an initial roster of relevant actors involved in peace processes. While it provides a comprehensive list of actors and groups, it lacks detailed information on the collaborative dynamics among these entities. On the other hand, the ICG's Crisis Watch provides a broader perspective by presenting a detailed timeline of conflict events, including mediation efforts and agreements. This timeline offers precise dates for events, enhancing the detail of our dataset and aiding in contextualizing the broader conflict resolution process.

3.2. Step 2: PA-X Contribution

Following the compilation of initial sources, PA-X Peace Agreements Database contributes by providing a list of mediation events related to signing agreements. PA-X includes peace agreements, ceasefires, and related documents, allowing researchers to trace back mediation events leading up to these agreements. While PA-X primarily focuses on events related to agreement affirmations, it too offers insights into the actors associated with these agreements, enriching our dataset with additional information on third-party involvement and collaboration dynamics.

3.3. Step 3: Reports from IGOs

In the third step of the data collection process, reports from international organizations (IOs) such as the United Nations (UN) are gathered to obtain a detailed timeline of events within conflicts. Reports submitted to the United Nations Secretary-General (UNSG) provide in-depth insights into UN-led peace processes, offering comprehensive coverage of ongoing mediation efforts. Additionally, reports from other IGOs such as the African Union (AU) and the Intergovernmental Authority on Development (IGAD) offer detailed insights into mediation processes within conflicts, albeit with varying levels of systematic reporting compared to UNSG reports. These reports contribute to our understanding of mediation events, identify additional actors and groups involved, and provide context for the broader peace process.

3.4. Step 4: Insights from Country Experts

The final preparatory step involves consulting with country experts to identify any actors, groups, committees, agreements, or other relevant information that may have been missed in the previous steps. Country experts provide valuable insight into local or third-party stakeholders and offer additional context that enhances the completeness of the dataset. By collaborating with experts familiar with the conflict context, researchers ensure the thoroughness and accuracy of the data collection process.

4. Semi-Automation: Enhancing Data Collection Efficiency and Depth

In the subsequent phase of our data gathering process, steps 5 through 6 transition into a semi-automated approach aimed at enhancing the comprehensiveness and validity of our dataset. This approach capitalizes on innovative methodologies and cutting-edge technologies to expand the scope of data collection while maintaining rigorous standards of accuracy and reliability.

4.1. Step 5: Comprehensive Web Crawling for Source Aggregation

To initiate the semi-automation process, we employ an advanced web crawling technique based on Kalev Leetaru's methodology. This approach leverages machine learning algorithms to aggregate sources related to third-party mediation activities. Unlike traditional manual methods such as LexisNexis and ReliefWeb, our web crawling technique operates on a vast scale, scouring over 1 billion references per year across global media outlets, digitized books, academic literature, human rights archives, and raw closed captioning streams television, among others.

The web crawling process identifies instances where there is a mention of mediation-like activities involving third-party actors within the conflict countries. This initial data pull generates millions of observations, capturing a diverse range of mediation events and actors operating in various contexts. However, due to the underspecified criteria[2] and the real-time translation of machine learning outputs into over 65 languages, the extracted data often contains noise and irrelevant information, necessitating detailed filtering in the subsequent steps.

To pinpoint mediation-type events within the universe of potential sources, this project employs Google Cloud's Computer Vision capabilities. This involves harnessing Optical Character Recognition (OCR) to detect text in raw files, swiftly summarizing them for rapid comprehension of vast textual data. Furthermore, it harnesses the power of Document AI, a document understanding platform adept at extracting text and data from scanned documents. This conversion of unstructured data into structured information streamlines the in-depth analysis of mediation-related content. Moreover, the Custom Classifier via Cloud Vision AI is integrated to discern potential mediation-related events. By identifying geographic locations and detecting the presence of third-party actors external to the specified content, this classifier enhances the precision of potential mediation categorization, enriching the understanding of mediation dynamics within the analyzed data, and ultimately providing a "universe" of potential sources before the filtering process.

4.2. Step 6: Refinement of Source List through Keyword Filtering

Following the comprehensive web crawling process, the next step involves refining the initial source list to facilitate the manual coding of the mediation event database. This refinement process unfolds in several stages, each aimed at enhancing the relevance and accuracy of the dataset.

First, we establish generic search criteria encompassing terms such as "mediation," "negotiation," "peace talks," "summit," "host talks," "peace process," "peace agreement," and

2 Implemented in order to ensure no observations are neglected.

"ceasefire." These predefined keywords serve as filters to eliminate irrelevant data noise resulting from underspecified search criteria or references to unrelated topics such as climate talks.

Additionally, leveraging our detailed list of actors, agreements, processes, and groups derived from earlier steps, we identify specific references to these terms within the sources. This targeted approach enables us to capture additional mediation events and actors that may have been overlooked in the manual data collection process.

By refining the initial source list through keyword filtering, we aim to enrich the dataset with a more comprehensive list of events associated with existing peace actors and processes. This iterative process of data refinement ensures the accuracy and relevance of the dataset, laying the groundwork for manual cross-checking and analysis in the subsequent steps.

This process narrows down the pool of cases, grouping similar events together and assigning keywords associated with their sources. Consequently, a comprehensive yet precise set of sources is presented to the manual coder, who decides whether they should be included in our final dataset and how their unique information should be coded. Providing this extensive array of sources to the coder helps avoid the pitfalls of previous manual searches, which often resulted in excessive and unnecessary searching. It also presents data from a larger universe than a manual coder could realistically navigate. Nonetheless, involving a manual coder at this stage ensures data accuracy.

Through the semi-automation process, GLO-PAD achieves a balance between both depth by pulling from a more complete universe of sources and efficiency by filtering and collapsing data to create a manageable yet comprehensive dataset, harnessing the power of technology to expand the scope of data collection while maintaining the integrity and reliability of the dataset. This innovative approach represents a significant leap forward in conflict mediation research, enabling researchers to access a wealth of data previously inaccessible through traditional manual methods alone.

5. Conclusion and Future Directions

The Global Peace Actors Database (GLO-PAD) has significant implications for advancing conflict resolution research by providing researchers with a comprehensive dataset for analysis. By capturing a broad spectrum of mediation events and actors, GLO-PAD enables scholars to identify patterns, assess the effectiveness of mediation efforts, and inform policy decisions.

This work contributes to conflict resolution research by providing a comprehensive dataset that expands the temporal and spatial scope of previous efforts, includes diverse mediation events, evaluates mediator bias, identifies new mediation actors, and utilizes innovative data collection

methods. In doing so, we aim to advance knowledge in the field of conflict resolution and inform more effective peacebuilding interventions.

GLO-PAD represents a valuable resource for researchers and practitioners in the field of conflict resolution. By employing a layered approach to data collection and leveraging semi-automated techniques, GLO-PAD offers insights into the complexities of mediation processes, thereby contributing to our understanding of peacebuilding dynamics.

Moving forward, GLO-PAD will continue to evolve, incorporating additional sources and refining its methodologies to enhance the accuracy and reliability of the dataset. Collaborative efforts among researchers and practitioners will be essential for maximizing the utility of GLO-PAD and advancing conflict resolution research.

References

International Crisis Group. (2023). Crisiswatch Database. Retrieved from https://www.crisisgroup.org/crisiswatch/database?location%5B%5D=5&date_range=custom&from_month=01&from_year=2021&to_month=12&to_year=2023

Leetaru, K. H. (2012). Fulltext geocoding versus spatial metadata for large text archives: Towards a geographically enriched Wikipedia. D-Lib Magazine, 18(9/10).

PA-X. (n.d.). PeaceRep: Peace Agreements Database. Retrieved from https://www.peaceagreements.org/

Universitat Autònoma de Barcelona. (n.d.). Peace Negotiations in Africa. Peace Talks in Focus. Report on Trends and Scenarios. Retrieved from http://escolapau.uab.cat/img/programas/alerta/negociaciones/23/africai.pdf

UNSG. (n.d.). UN Documents for Central African Republic: Secretary-General's Reports. Security Council Report. Retrieved from https://www.securitycouncilreport.org/un_documents_type/secretary-generals-reports/?ctype=Central+African+Republic&cbtype=central-african-republic

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17826

Data-Driven Strategies for Early Detection of Corporates' Financial Distress

Donato Riccio[1], Giuseppe Bifulco[2], Paolone Francesco[3], Andrea Mazzitelli[4], Fabrizio Maturo[4]

[1]Machine Learning Engineer, Student at the Master's in Data Science at the University of Campania Luigi Vanvitelli, Caserta, Italy, [2]Department of Economics, Management, Institutions, University of Naples Federico II, Naples, Italy, [3]Faculty of Economic and Legal Sciences, Universitas Mercatorum, Rome, Italy, [4]Faculty of Technological and Innovation Sciences, Universitas Mercatorum, Rome, Italy.

How to cite: Riccio, D.; Bifulco, G.; Paolone, F.; Mazzitelli, A.; Maturo, F. 2024. Data-Driven Strategies for Early Detection of Corporates' Financial Distress. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17826

Abstract

Scholars have taken a keen interest in predicting corporate crises in the past decades. However, most studies focused on classical parametric models that, by their nature, can consider few predictors and interactions and must respect numerous assumptions. Over the past few years, the economy has faced a severe structural crisis that has resulted in significantly lower income, cash, and capital levels than in the past. This crisis has led to insolvency and bankruptcy in many cases. Hence, there is a renewed interest in research for new models for forecasting business crises using novel advanced statistical learning techniques. The study shows that using tree-based methods and hyper-parameters optimization leads to excellent results in terms of accuracy. This approach allows us to automatically consider all possible interactions and discover relevant aspects never considered in past studies. Furthermore, we employ SHAP (SHapley Additive exPlanations) to enhance the explainaibility of our model. This line of research provides fascinating results that can bring new knowledge into the reference literature."

Keywords: *corporate crises; financial distress; statistical learning*

1. Introduction

The first forecasting models to diagnose the state of corporate health date back to the 1960s and 1970s. These models were based on balance sheet indexes, which are still valid tools for preventing the company's state of health. The most used in the literature were profitability (ROE, ROI, ROS, Turnover), liquidity, and capital-financial solidity (e.g. debt ratio). The reference economic and business doctrine furnishes numerous contributions to developing forecasting

models based on the combined observation of performance indicators (Altman, 1968; Altman et al., 2013; Altman and Hotchkiss, 2006; Jones and Hensher, 2004; Shumway, 2001).

In recent years, the economy has gone through a deep, sometimes irreversible, structural crisis, where the settling of income, cash, and capital levels has been significantly lower than in the past. In many circumstances, this crisis has led to insolvency and bankruptcy. Therefore, there is an urgent need to determine warning signs to promptly and effectively activate a recovery process before business continuity is compromised. To this end, there is a growing interest in combining quantitative models based on refined statistical learning techniques that, considering large volumes of data, contemplate the temporal aspect of the balance sheet data, the dynamic element of the context variables, and the interactions of these data over time. Accordingly, there is a revitalised attraction in research for new models for forecasting business crises using novel advanced statistical learning techniques. The study reveals that using XGBoost (Chen, 2016) and hyper-parameter optimisation leads to superior predictive power results. While ensemble methods are often considered black-box models, we employ SHAP (SHapley Additive exPlanations) (Lundberg, 2017) to interpret their predictions, ensuring accuracy and explainaibility.

There are often large datasets available regarding sample size and the number of variables in the business field. However, having many statistical units is undoubtedly a great advantage when the goal is to create a classifier but, on the other hand, having a large number of variables available leads to the so-called curse of dimensionality, which leads to different methodological problems such as the choice of the model, the sparsity of data, the concentration of distances and, above all, multicollinearity. All these aspects make classical parametric approaches obsolete.

In statistics, tree-based classifiers have numerous advantages: they overcome the problems related to the curse of dimensionality, significantly improve performance both in terms of precision and reducing the estimates' variability, offer a dynamic interpretative key to the determinants of a phenomenon, and do not require particular assumptions to be respected. Moreover, tree-based classifiers automatically consider all possible interactions and uncover relevant aspects that may have yet to be considered.

Our study shows that integrating the latest machine learning techniques in this area significantly benefits prediction and explainability. Since these techniques have been little used in the business field to predict the state of crisis, further studies are recommended, introducing context variables and spatial analysis.

2. Material and Methods

The data used for the study were collected from the AIDA dataset, which is a database created and distributed by Bureau van Dijk S.p.A. It contains the balance sheets, personal data, and product information of active and failed Italian capital companies, excluding banks, insurance companies, and public bodies.

The sample selection criteria are based on Legislative Decree 139/2015 and aim to identify Italian non-financial companies that can be categorised as medium or large based on at least two of three thresholds. These criteria exclude small businesses, which are defined as companies that do not exceed two of the following limits at the balance sheet closing date: a net equity total of 4 million EUR, net sales revenue of 8 million EUR, and an average number of 50 employees during the fiscal year. The final sample size of medium to large Italian non-financial companies that meet these specified benchmarks is 37,369.

The research utilises a comprehensive set of financial variables from 2015, 2016, and 2017 to analyse Italian non-financial companies. These variables include annual sales revenue in millions of EUR, EBITDA in millions of EUR, net profit in millions of EUR, total assets in millions of EUR, and the number of employees. Ratios such as EBITDA to sales percentage, debt to EBITDA percentage, and invested capital turnover are included to assess profitability and financial efficiency. The analysis also considers the net equity in millions of EUR, short and long-term debt ratios, liquidity ratios, current ratios, coverage indices of immovable assets, net financial position in millions of EUR, and debt-to-equity ratios. Furthermore, it evaluates financial autonomy from third parties, financial charges on turnover, interest coverage, the cost of borrowed money, bank debts to turnover, and various profitability ratios like Return on Equity (ROE), Return on Sales (ROS), Return on Investment (ROI), and Return on Assets (ROA). Tax liabilities, both short-term and long-term, are also included. Additional variables account for differences between years for these metrics, capturing changes over time. Geographical variables are included based on the province (e.g., Brescia, Milano, Roma, Torino, other) and the sector of activity coded according to NACE (e.g., 25, 28, 41, etc.), with additional categories for other sectors. These variables enable a detailed analysis of the financial health, performance, and geographical and sectoral distribution of the companies in the sample. Considering the differences between years, there are a total of 181 variables in the dataset. The dataset, consisting of 434 instances, was divided into a training set containing 347 instances (80%) and a test set with 87 instances (20%). The training and test sets were balanced regarding the target variable.

The research employs an undersampling technique on the majority class, stratified according to the NACE sector codes. This approach ensures a balanced representation across different sectors by reducing the size of the majority class to match that of the minority class. After the undersampling process, the dataset achieved equilibrium with 217 observations for each class,

labelled '0' and '1', resulting in a balanced dataset. The dataset, consisting of 434 instances, was divided into a training set containing 347 instances (80%) and a test set with 87 instances (20%). The training and test sets were balanced regarding the target variable.

We employed XGBoost (Chen, 2016), a robust machine learning algorithm that uses gradient boosting frameworks for predictive modelling, which is renowned for its performance and speed in classification tasks. The target variable is labeled 1 if the company started filing a procedure for bankruptcy in 2018 or 2019, and 0 otherwise.

For hyperparameter tuning, we utilized Bayesian optimization (Bergstra, 2015), a probabilistic model-based approach for global optimization, running 100 evaluations to efficiently converge on the optimal set of parameters. To validate our model, we implemented a 10-fold cross-validation technique.

3. Results and Conclusions

Without hyperparameter tuning, the initial model achieved a mean ROC AUC score of 0.8990 using 10-fold cross-validation, demonstrating the algorithm's strong predictive capabilities out-of-the-box. After conducting thorough hyperparameter tuning to optimise the model's performance, we obtained an improved mean ROC AUC score of 0.9199, indicating a substantial enhancement in the model's ability to discriminate between bankrupt and non-bankrupt companies. Furthermore, the test set AUC score is 0.9413, ensuring good predicting performance on unseen data.

The SHAP summary plot in Figure 1 visualizes the impact of various financial metrics on the model's prediction across many observations. Positive SHAP values (red) indicate features that positively influence the model's outcome, whereas negative values (blue) show a decreasing effect. Key influencers include net income, return on equity, and the interest coverage ratio, which vary across different data points, showing both strong positive and negative effects on the model's output, highlighting their critical role in determining financial health and stability.

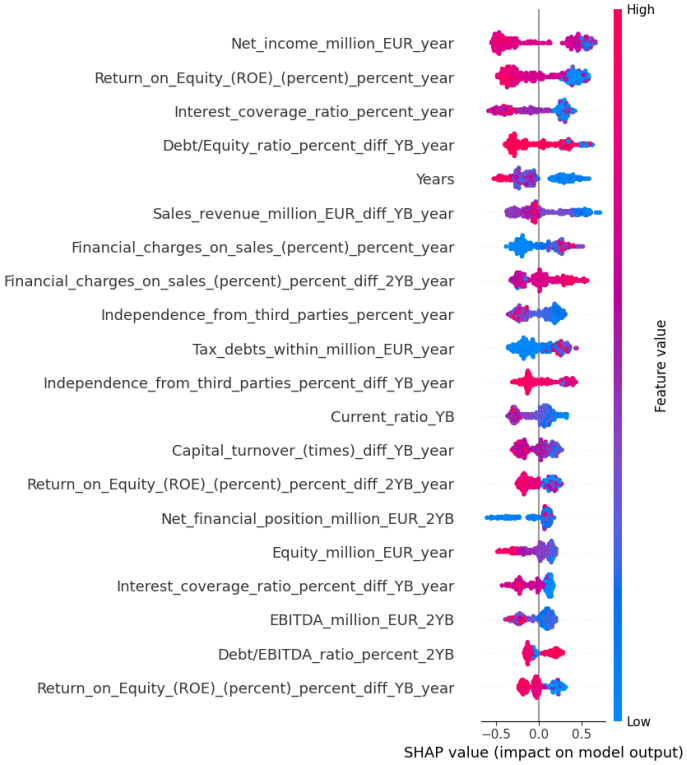


Figure 1. Global explanation SHAP plot.

Figure 2 and Figure 3 illustrate how various financial indicators influence the predicted probability of a company's failure. For a financially healthy company. Features like high net income, positive changes in interest coverage ratio, and return on equity significantly decrease the risk of failure. A large negative financial position increases the likelihood of failure, demonstrating the impact of financial health on company stability. Notably, a negative ROE raises the probability. A negative net income and increases in tax debts within the year also contribute positively.

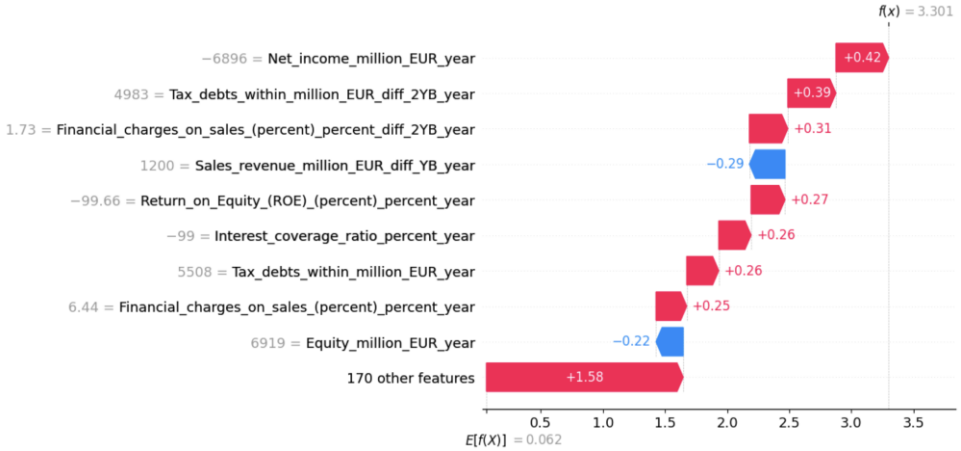


Figure 2. Local Explanation for a Financially Distressed Company.

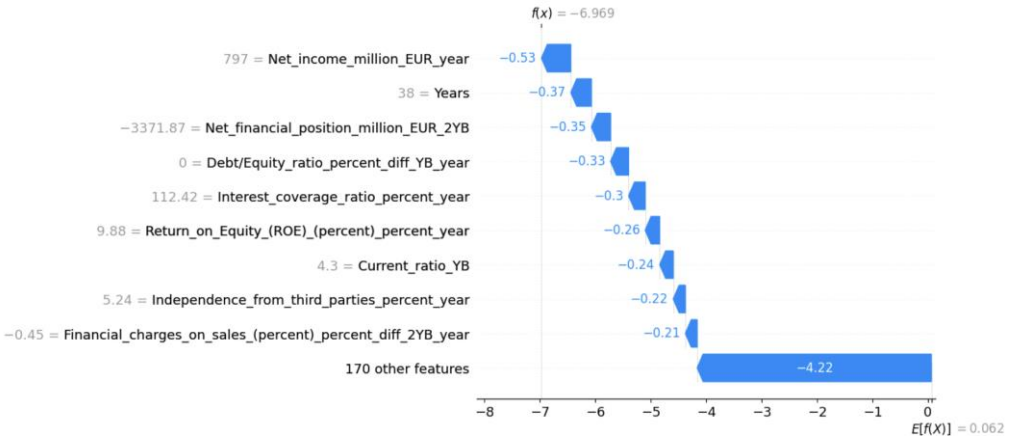


Figure 3. Local SHAP Explaination for a financially healthy company.

The research findings reveal the significant potential of advanced machine learning techniques, particularly tree-based ensemble methods like XGBoost, in predicting corporate financial distress. By leveraging a comprehensive set of financial variables and employing rigorous hyperparameter optimisation, our model achieves superior performance compared to traditional parametric approaches. Moreover, by integrating SHAP, our work ensures that our model maintains explainability, allowing for a clear understanding of the key factors driving the predictions. The SHAP summary plot and local explanations provide valuable insights into the impact of various financial indicators on a company's risk of failure. This combined approach improves predictive performance and offers an explainable approach to corporate crisis forecasting, bridging the gap between accuracy and interpretability. This line of investigation promises compelling insights that have the potential to enrich the current literature. Further

insights and developments on the topic will provide exciting results in generalising to more countries and sectors.

Funding

This research is funded as part of a project supported by the Mercatorum University of Rome under grant code *12-FIN/RIC 2023*.

References

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. Journal of Finance, 23(4), 589-609.

Altman, E. I., Danovi, R., & Falini, A. (2013). Z-Score models' application to Italian companies subject to extraordinary administration. Bancaria, 4, 24-37.

Altman, E. I., & Hotchkiss, E. (2006). Corporate Financial Distress and Bankruptcy (3rd ed.). John Wiley & Sons.

Jones, S., & Hensher, D. A. (2004). Predicting firm financial distress: A mixed logit model. Journal of Accounting and Public Policy, 23(6), 467-487.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. Journal of Business, 74(1), 101-124.

Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765-4774).

Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., and Cox, D. D. (2015). Hyperopt: a Python library for model selection and hyperparameter optimization. Computational Science & Discovery, Volume 8, Number 1.

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17405

Multilingual Monetary Policy: Unfolding Language and Policy Preferences of Swiss Central Bankers

Sami Diaf[1], Florian Schütze[2]

[1]Department of Socioeconomics, University of Hamburg, Germany, [2]Faculty of Economics and Social Sciences, Helmut Schmidt University, Germany.

How to cite: Diaf, S.; Schütze, F. 2024. Multilingual Monetary Policy: Unfolding Language and Policy Preferences of Swiss Central Bankers. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17405

Abstract

Understanding monetary policy has always been of paramount economic and political importance. However, it remains a difficult task, despite transparency efforts and the regular flow of information to the public, which becomes even more complex when communication channels are multilingual. This paper examines the policy narratives of the Swiss National Bank (SNB) in terms of language and policy preferences, using the corpus of speeches delivered by its members over the period 1997-2022. Using a dynamic semantic search strategy based on top2vec, the framework analysis was able to identify interlingual similarities and differences with the help of pre-trained multilingual models. The results show that the SNB's communication strategy is strongly oriented towards the objectives assigned to the central bank, with attention being paid to systemic risks, banking regulation and financial markets, which emerge as second but no less important objectives, closely linked to the international environment, in particular the Eurosystem as a strategic aspect of the stability of the Swiss franc. The results suggest that English is used exclusively to address core central banking issues (monetary policy, inflation and interest rates), while uncertainty concerns seem to be reported more in German or French. The resulting dual semantic space, consisting of embedded words and documents, yielded relevant topics with respect to the size and scope of the corpus. Furthermore, informative indices could be constructed for policy measurement, as a crisis index was found to be consistent with the business cycle fluctuations and technical recessions experienced in Switzerland over the last 25 years.

Keywords: *Topic model; Natural Language Processing; Web scraping ; Explainability and interpretability; Finance applications.*

1. Introduction

Narrative economics has established itself as a growing field of research in economics (Shiller, 2017), benefiting from the continuous development of technical solutions that allow powerful analysis of documents (Grimmer et al., 2022). They mostly use topic models as a popular class of machine learning algorithms dedicated to text data (Gentzkow et al., 2019).

Such unsupervised models have been the workhorse of text mining applications, whose goal is to uncover latent groups of words, known as topics, from the triplet structure document-topic-word (Blei et al., 2003), which provides a deeper understanding of the corpus, although it remains exclusively used at the monolingual level. While a growing number of corpora are available in different languages, they require advanced mechanisms to ensure interlingual translation in topic discovery (Lucas et al., 2015).

Recent advances in Natural Language Processing (NLP) provided ways to handle multilingual documents in the same corpus (Bianchi et al., 2021) by exploiting the power of distributional representations (Mikolov et al., 2013; Dieng et al., 2019), such that words and paragraphs can be vectorized to ensure a better representation of topics through semantic search models (Angelov, 2020; Grootendorst, 2022).

This work adopts the latter approach to shed light on the historical evolution of the SNB's communication over the last 25 years by analyzing the trilingual corpus of speeches of its members (1997-2022) using semantic search models, namely top2vec (Angelov, 2020), which proved to be useful for obtaining powerful and concise topic representations for both monolingual and multilingual corpora. In particular, the resulting semantic space yields robust topic features and allows testing further hypotheses about the rhetoric central bankers often use when addressing their audience, as well as quantifying other interests related to uncertainty and crisis, two frequently discussed topics in modern macroeconomics.

The results suggest that SNB communication is focused on the objectives assigned to central bankers, as the number of topics (8) learned by top2vec is relatively low compared to the depth of the corpus (669 documents). In addition, secondary objectives such as systemic risk, banking supervision and financial stability were regularly discussed since the outbreak of the financial crisis in 2008, while developments in the euro area remained relevant in the discourse as an anchor of stability for the Swiss franc even after the abandonment of the euro peg. Finally, topics related to standard monetary policy practices, such as unemployment and inflation, were mostly reported in English, in contrast to uncertainty, which tended to appear in German and French speeches.

2. Central Banking Communication

Monetary policy, as a dynamic subfield of economics, has witnessed a growing number of applications with textual data, in particular aimed at extracting sentiments from speeches and minutes. This is done to study their impact on financial markets (Grimmer & Stewart, 2013) or to uncover the policy preferences of monetary policy committees (Hansen et al., 2018; Diaf, 2022; Shapiro & Wilson, 2022; Diaf & Schütze, 2023), mostly targeting monolingual corpora, where the methods used are derived from Latent Dirichlet Allocations (Blei et al., 2003).

Central banks around the world have used a variety of channels to maintain a permanent informative link with the public. Many of them have adopted a multilingual communication strategy to reach a wider audience (Athanassiou, 2006; ECB, 2011), especially for multicultural countries or monetary unions. In this way, central banks disseminate the same information in many languages to avoid any bias resulting from translation or the use of a coded language (Muchlinski, 2011, pp. 224–225), a strategy that increases the public's perception of uncertainty. Common references in this area are the European Central Bank (ECB), the Deutsche Bundesbank and the Swiss National Bank (SNB), whose communication operates in three main languages, making it a good example of how multilingualism works as an adopted communication strategy.

For this purpose, it is necessary to use a framework that unifies different languages to ensure a fair assessment of the signals emerging from the corpus and to avoid potential biases that arise when adopting context-dependent, agnostic translation schemes. Pre-trained multilingual models can be used efficiently for such tasks and have proven to be highly informative for multilingual central banking corpora (Diaf & Fritsche, 2022).

3. Application

The corpus of SNB speeches was scraped from the SNB website and consists of 669 different speeches given between 1997 and 2022 in three languages: English (374), French (90), and German (205). They were given as input without any preprocessing steps to top2vec, which was augmented with a pre-trained multilingual sentence transformer embedding model to unify the framework and avoid setting as many models as languages that require a post-hoc translation mechanism (Lucas et al., 2015). Table 1 shows the resulting topics (8) learned using *top2vec* after mapping sentences and paragraphs to a 512-dimensional dense vector space, which is used for tasks such as clustering or semantic search.

Table 1: Top 20 words of the topics learned using top2vec. Source: Own calculation.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
wahrungspolitik	bankensystem	wahrungspolitik	inflationsraten	euroraum	nationalbank	wahrungspolitik	gold
inflationsraten	bankensystems	finanzmarkten	inflationsrate	euro	geschaftsbanken	devisenmarkten	wahrungspolitik
inflationsrate	finanzkrise	finanzmarkte	inflation	eurogebiet	kantonalbanken	geldmarktzinsen	geldmarktzinsen
inflationary	geschaftsbanken	inflationsraten	inflationary	eurogebiets	bankensystem	geldmarkt	wahrungsfonds
inflation	finanzsystem	geldmarktzinsen	inflationnistes	eurozone	banknotes	devisenmarkt	banknote
inflationnistes	finanzsystems	finanzmarkt	wahrungspolitik	euros	bankensystems	finanzmarkten	banknotes
geldpolitik	zentralbanken	inflationsrate	economist	wahrungspolitik	zentralbanken	currencies	nationalbank
finanzpolitik	finanzmarkten	wahrungsfonds	economists	currencies	banknote	finanzmarkte	monetary
finanzkrise	zentralbank	currencies	inflationsziel	euroland	zentralbank	finanzmarkt	monetaire
economist	finanzmarkte	geldmarkt	economique	europaischen	banknoten	currency	geldpolitik
economists	bundesbank	finanzkrise	economiques	european	bundesbank	inflationsraten	kantonalbanken
geldpolitischen	bankensektor	devisenmarkten	finanzkrise	europas	wahrungspolitik	inflationsrate	monetaren
geldpolitische	finanzmarkt	finanzpolitik	volkswirtschaft	europaische	bankensektor	wahrungen	banknoten
inflationsziel	kantonalbanken	inflationary	economy	europe	bank	monetaire	geldpolitische
geldpolitisches	nationalbank	banknotes	economics	inflationsraten	interbank	inflationary	geldpolitisches
economique	grossbanken	inflation	macroeconomic	currency	grossbanken	monetary	geldmarkt
geldpolitisch	interbank	inflationnistes	recession	inflationsrate	banken	monetare	geldpolitischen
fiskalpolitik	banks	devisenmarkt	wirtschaftliche	inflationary	banque	inflation	bankensystem
geldpolitischer	bank	geldpolitik	economie	europeen	bancaires	monetaren	monetare
economiques	bancaires	economists	economies	europeenne	banking	inflationnistes	geschaftsbanken

The topics listed in Table 1 in order of importance have an overlapping structure and are closely related to the objectives assigned to the SNB, such as ensuring price stability (topics 1 and 4), financial market stability (topics 3 and 7), the banking system (topics 2 and 6) and the currency (topics 5 and 8). Topic 1 could be identified as the practice of monetary policy, which is closely related to fiscal policy (policy mix), while topic 4 specializes in inflation targeting.

Table 2: Top 20 terms related to ”uncertainty”, and their correlation given by the cosine similarity of their respective vectors. Source: Own calculation.

Term	Corr.	Term	Corr.
uncertainties	0.982	indebtedness	0.770
incertitude	0.971	decisively	0.769
incertitudes	0.946	ungefahr	0.758
uncertain	0.931	preisstabilitaet	0.744
unsicherheiten	0.888	glaubwurdigkeit	0.744
unsicherheit	0.887	inevitable	0.744
unabhangigkeit	0.806	instability	0.742
doubts	0.802	unerwunschten	0.740
unsecured	0.791	conviction	0.737
doubt	0.771	reliability	0.734

Table 3: Top 20 terms related to ”crisis”, and their correlation given by the cosine similarity of their respective vectors. Source: Own calculation.

Term	Corr.	Term	Corr.
crise	0.983	konjunkturellen	0.692
krise	0.979	jeopardise	0.686
krisen	0.978	deflation	0.685
crises	0.963	deflationary	0.664
konjunktur	0.784	threatened	0.659
finanzkrise	0.783	risikopraemie	0.658
recession	0.738	droht	0.655
konjunkturlage	0.721	collapse	0.649
rezession	0.714	ausbruch	0.645
konjunkturelle	0.705	einbruch	0.642

The multiple occurrences of words in different topics indicate different contexts used by central bankers in their speeches, depending on the prevailing macroeconomic and financial situation and the "hot topics" they were trying to address. For example, Topic 4 deals with inflation in a

macroeconomic context, while Topic 1 still discusses inflation from an inflation targeting perspective, close to fiscal policy, suggesting a latent policy mix debate. The topic content uncovered from the speeches over 25 years is limited by a relatively small number of terms compared to the 10 different topics learned with the same methodology using Bundesbank speeches over the period 2012-2017 (Diaf & Fritsche, 2022), which may indicate a potential ambiguity (Baerg, 2020) used at the SNB.

Tables 2 and 3 provide specific details on the use of words related to uncertainty and crisis within the SNB. Although central bank jargon is rich in technical terms related to macroeconomics and financial markets, the use of uncertainty seems to be limited to its literal aspect (in the three languages). Crisis-related terms mostly have a macroeconomic connotation related to the state of the economy and the financial/banking system.

The distances of each document to the word "crisis", in terms of the cosine similarities of their respective vectors, were used to construct a crisis index to measure its prevalence within each document in the corpus. Although the documents are irregular in terms of publication, a monthly aggregated index was constructed and found to correlate with business cycle fluctuations in Switzerland (Figure 1). The use of crisis-related rhetoric seems to color SNB speeches during technical recessions, except for the period 2011-2013, which coincides with the euro peg adopted by the SNB.

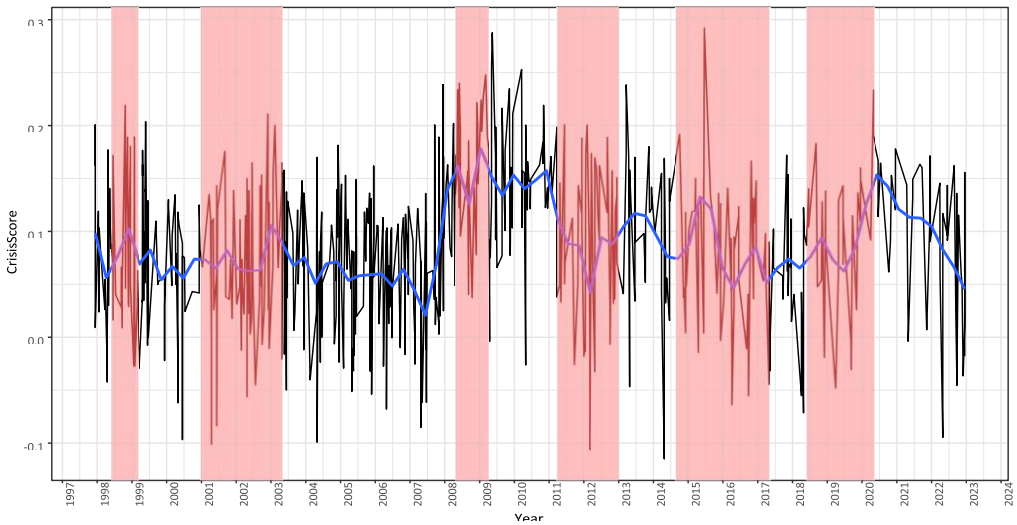


Figure 1: Crisis-scores per document and recession periods (in shaded pink). Blue curve represents the LOESS smoothed curve. Source: Own calculation.

4. Conclusion

Central bank communication has long been a black box when it comes to understanding monetary policy signals delivered through their communication channels, and this difficulty increases when the strategy adopts multilingualism to achieve greater transparency in their missions. Recent advances in natural language processing make it possible to unify the analysis framework of multilingual corpora, typically using the power of distributional representations and semantic search architectures that uncover meaningful word associations and robust topic structures, instead of classical bag-of-word techniques that often require translation mechanisms and biased human intervention. The multilingual corpus of SNB speeches over the past 25 years reveals a communication strategy that is essentially built around the goals assigned to the SNB, with a particular interest in financial markets, the banking system, and currency stability. These developments are important in guiding modern monetary policy practice; however, the corpus does not offer practical discussions or specific jargon that might color the topics compared to other corpora from Eurosystem central banks. The reduced number of topics (8) compared to the depth and breadth of the corpus does not unfold granular structures that could be assimilated to subtopics, indicating a moderately informative communication style. This is often used to secure financial markets and is potentially colored with ambiguity as a stylized fact of modern central bank communication. Beyond the topic results, semantic search strategies help to build useful associations within the document-topic-word triplet, such as for crisis-related terms, which proved to highlight central bankers' responsiveness to turmoil and business cycle fluctuations, as well as demonstrating a clear preference for the English language when dealing with standard monetary policy concerns.

References

Angelov, D. (2020). Top2vec: Distributed representations of topics. arXiv preprint arXiv:2008.09470.

Athanassiou, P. (2006). The application of multilingualism in the european union context. ECB Legal Working Paper No. 2.

Baerg, N. (2020). Crafting consensus: Why central bankers change their speech and how speech changes the economy. New York, NY, United States of America: Oxford University Press.

Bianchi, F., Terragni, S., & Hovy, D. (2021). Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 759–766.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), pp. 993–1022.

Diaf, S. (2022). Policy preference at central banks: Quantifying monetary policy signals using keyword topic models. WiSo-HH working paper series 69.

Diaf, S. & Fritsche, U. (2022). TopicShoal: Scaling partisanship using semantic search. In Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022), Potsdam, Germany, pp. 167–174.

Diaf, S., Schütze, F. (2023), Estimating Policy Uncertainty Within Monetary Policy Debates, CARMA 2023 - 5th International Conference on Advanced Research Methods and Analytics, Sevilla, Spain, pp. 269–277, DOI: 10.4995/CARMA2023.2023.16419

Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2019). Topic modeling in embedding spaces. arXiv preprint arXiv:1907.04907.

ECB (2011). The monetary policy of the ECB. Frankfurt am Main, Germany, ECB, ISBN 978-92-899-0777-4

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. Journal of Economic Literature, 57(3), 535–574.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). Text as Data: A New Framework for Machine Learning and the Social Sciences. United States of America: Princeton University Press.

Grimmer, J. & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political Analysis, 21(3), 267–297.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.

Hansen, S., McMahon, M., & Prat, A. (2018). Transparency and deliberation within the FOMC: A computational linguistics approach. The Quarterly Journal of Economics, 133(2), pp. 801–870.

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. Political Analysis, 23(2), pp. 254–277.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Muchlinski, E. (2011). Central Banks and Coded Language: Risks and Benefits. London, UK: Palgrave Macmillan.

Shapiro, A. H. & Wilson, D. J. (2022). Taking the fed at its word: A new approach to estimating central bank objectives using text analysis. The Review of Economic Studies, 89(5), pp. 2768-2805.

Shiller, R. J. (2017). Narrative economics. The American Economic Review, 107(4), pp. 967–1004.

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17554

Unlocking the Potential of Machine Learning in Portfolio Selection: A Hybrid Approach with Genetic Optimization

Chaher Alzaman

Department of Supply Chain and Business Technology Management, John Molson School of Business, Concordia University, Montreal, Quebec, H3H 0A1, Canada.

How to cite: Alzaman, C. 2024. Contemporary issues in Financial Technology: the role of the Internet. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17554

Abstract

In the field of financial market predictions, machine learning has been widely used to identify patterns and gain valuable insights. However, for success in portfolio selection, it is crucial to optimize factors that impact accuracy. This study focuses on combining machine learning and optimization to enhance stock selection and prediction capabilities, thereby addressing a critical challenge faced by investors and traders. The work starts with hyper-parameter optimization and utilizes three different machine learning algorithms: XGBoost, LSTM, and Deep RankNet. These algorithms were chosen for their proven performance in handling complex financial data and capturing nonlinear relationships. Our findings show a 40% improvement in results through the use of a genetic-based optimization technique, as well as a promising daily average return of 0.47% through a novel feature engineering approach. The study provides a framework for optimizing and learning in financial portfolio selection, with promising results for medium and small-sized traders who often face resource constraints in developing sophisticated trading strategies. The proposed approach offers a scalable and adaptable solution that can be tailored to different market conditions and investment objectives.

Keywords: *Artificial Intelligence; Machine Learning; Optimization; Financial Markets; Predictive Analytics.*

1. Introduction

Machine learning has gained widespread attention in the financial market as a tool for predicting stock prices, foreign exchange rates and other market trends. With its ability to analyze large amounts of data, machine learning algorithms can provide more accurate predictions compared to traditional statistical methods. Shah (2007) highlights two main approaches in stock

prediction: Fundamental Analysis, where analysis is based on a company's financial characteristics (such as past performance, assets, earnings, etc.), and Technical Analysis, where patterns in past stock prices are studied. Despite the efficient market hypothesis (Jensen, 1978) stating that stock prices cannot be predicted and the random-walk hypothesis (Malkiel, 1973) suggesting stock prices only depend on future information and not on history, research by Basak et al. (2019), Chen et al. (2020), and others argue that some elements of stock behavior are predictable.

This work employs LSTM, XGBoost, and Deep RankNet. To set a background, two classes of methods have been prominent in the literature (Machine Learning applications in financial markets): Artificial Neural Networks (ANN) and Ensemble tree-based algorithms. ANN is at the heart of Deep Learning, which in turn is a subset of Machine Learning (ML) geared toward more complex systems (e.g., big data). LSTM (long short-term memory) is an artificial recurrent neural network tailored to sequential data, such as closing prices of financial assets. The XGBoost (XGB) is an ensemble decision tree-based algorithm that is quite popular in financial market predictions. Basak et al. (2020) and Chen et al. (2021) confirm the effectiveness of XGB. Our newest method, Deep RankNet, is a novel approach in predicting financial assets, first introduced by Burges et al. (2005). The concept behind Deep RankNet involves using deep learning to rank and match objects. Li and Tan (2021) have applied Deep RankNet to ranking and forming trading portfolios in the field of financial market predictions. This work supplements the above with Hyperparameter optimization, which involves adjusting critical parameters in machine learning. The user sets these parameters prior to training. But for complex and intricate systems such as financial systems, proper tuning is crucial as financial assets are notorious for being noisy and sensitive to small variations in the input.

2. Literature Review

Many studies in the field of predictive analytics in financial markets compare the effectiveness and performance of various Machine Learning (ML) techniques. These studies evaluate different ML algorithms and choose the best one(s) for financial market prediction. Some examples of these studies include works by (Basak et al., 2019, Kumar et al. 2018, Patel et al. 2019, Ismail et al. 2020, Usmani et al. 2016, Nelson et al. 2017, Shen et al. 2012, Singh and Srivastava 2017, and Vijh et al. 2020). The works mainly examine different strategies for predicting financial market outcomes. On the other hand, there are studies that aim to improve the prediction performance of one or more ML techniques, such as (Porshnev et al. 2013, Wang et al. 2018, Kim et al. 2020, Akita et al. 2016, Reddy 2018 and Vazirani et al. 2020).

In the field of Portfolio Selection, there have been several studies that apply different techniques. Chen et al. (2020) use XGBoost with multi-features, including technical indicators, for selection and hyperparameter optimization in the Shanghai Stock Exchange. Liu and Yeh (2017) employ

neural networks to predict the behavior of American stocks, but do not perform feature selection or hyperparameter optimization. Paiva et al. (2019) suggest using a Support Vector Machine (SVM) for stock predictions in Sao Paulo and conduct hyperparameter optimization. Long et al. (2019) utilize a multi-filter neural network for feature extraction and price movement prediction of financial time series data, using 1-minute stock price frequency and convolutional and recurrent neurons. It is important to note that small and medium investors may not have access to such detailed data, especially in smaller trading markets.

One of this work's main contributions is the explicit use of a genetic-based algorithm to tune the machine learning hyper-parameters. Dessain (2022) classifies predicting returns' performance metrics, within the context of machine learning, into error-based, accuracy-based, and investment-based metrics. The first metric is associated with computing prediction errors. The second metric is the measure of returns based on the algorithm's prediction accuracy. The last is concerned with result-based metrics and risk-adjusted return-based metrics. This work employs all three categories and focuses on risk-adjusted return-based metrics.

3. Methods

3.1. LSTM

LSTM is a form of a recurrent neural network. Figure 1 illustrates a schematic for LSTM, where multiple gates encompass activation functions. It closely follows the work of Fischer and Krauss (2018). LSTM can be thought of as a system of cells where information gets added or removed to the cell state by the use of gates depending on its importance (important information kept, irrelevant information discarded).

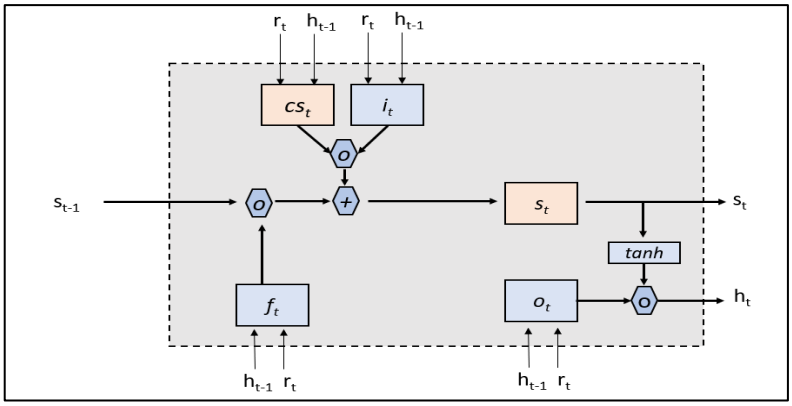


Figure 1: LSTM Architecture of memory cell (f, forget; i, input cell; cs, candidate state value; s, cell state; o, output cell)

We utilize 42-time steps (t), where at every t the input is posed as xt and an output ht-1 of the memory cell at the previous t (t-1). As per the work of Fischer and Krauss (2018), we utilize the stock return as an input parameter based on the given stock closing P (Pt) for a specific period t.

$$R_t = \frac{P_t}{P_{t-1}} - 1 \tag{1}$$

In essence, the gates act as filters (see figure 1). The forget gate dictates which information to remove from the memory cell state. While the input gate ushers which information to add to the memory cell state. Finally, the output gate directs which information from the memory cell state to utilize as output. The equations below characterize the LSTM procedure in a vector form. Candidate state shall be represented as cs, while state cell is indicated by s alone. Input and output values are designated as i and o.

- rt is an input vector at time step (t)
- The weight matrices are Wf,r, Wf,h, Wcs,r, Wcs,h, Win,h, Wo,r, and Wo,h.
- The bias vectors are bf, bcs, bin, and bo.
- Activation function vectors are ft, it, and ot for the three gates respectively.
- The output vector for the LSTM layer is ht.

The cell states and output are updated using the following procedure. First, LSTM needs to decide which information to discard. This is done using the activation function (Eqn. 2).

$$f_t = sigmoid(W_{f,r}r_t + W_{f,h}h_{t-1} + b_f) \tag{2}$$

We have ran two instances of a sigmoid and linear activation function, and found the first to be superior. Second, the LSTM layer decides which information to be added to the cell states in two parts. First, the candidate state value is computed (Eqn. 3). Second, the activation value of the input gate is computed (Eqn. 4)

$$cs_t = tanh(W_{cs,r}r_t + W_{cs,h}h_{t-1} + b_{cs}) \tag{3}$$

$$i_t = sigmoid(W_{i,r}r_t + W_{i,h}h_{t-1} + b_i) \tag{4}$$

In the third step, the new cell states s, t, are calculated based on the results of the previous two steps denoting the Hadamard (elementwise) product. Then, the Hadamard product is used to arrive at the new cell state value (Eqn. 5).

$$s_t = f_t \cdot s_{t-1} + i_t \cdot cs_t \tag{5}$$

Finally, the output of the memory cell is computed using equations 6 and 7.

$$o_t = sigmoid(W_{o,r} r_t + W_{o,h} h_{t-1} + b_o) \tag{6}$$

$$h_t = o_t \cdot \tanh(s_t) \tag{7}$$

The data input for the LSTM and all three algorithms constitutes thirty-two training and two holding days as per the work of Li and Tan (2021). However, the input for all three algorithms is amalgamated to enhance the learning capabilities of both algorithms. This is a primary contribution of this work. In effect, for each stock entry, six lags are created. So, the input matrix is the vector of S ($\forall s \in S \times NL$) stocks where NL stands for the number of lags applied. As shown in the exemplary matrix below, for eleven-day training data, we construct six lags where each lag contains six days vector.

$$\begin{bmatrix} R_1 & \cdots & \ldots & R_6 \\ R_2 & \ldots & \ldots & R_7 \\ \vdots & \ddots & \ddots & \vdots \\ R_5 & \cdots & \ldots & R_{10} \\ R_6 & \ldots & \ldots & R_{11} \end{bmatrix}$$

We have contested different NL values of {2, 3, 5, 6, 7, 8, 9} and found six to be optimal. Higher NL's renders a higher model complexity and a deterioration in performance. Second, we study different periods (size and boundaries) and subperiods as should be indicated in the results' figures in the next sections. Figure 2 exhibits the procedural code for the LSTM algorithm used in this work. A set of 100 TSE stocks S {s1, s2, …., s100} is used to compute daily returns (eqn. 1). Then the returns are ranked with respect to the median (Fischer & Krauss 2018), MS, of the set S. Stocks with higher returns than MS are assigned to Class 1, lower stocks are assigned to Class 0. The values of the hyperparameters are pre-set to the output optimal values of the Genetic Algorithm (to be discussed at the end of this section). Ultimately, the LSTM will handle the input matrix and output value of one or zeros depending on the class.

→Compute excess return for all ***Stocks***
→Compute the median of each stock
→Assign top 50% stocks that are higher than the median to *Class 1*
→Assign lowest 50% stocks a ranking label of *Class 0*
→Construct Input Feature Matrix of (N_{stocks} * NL)× N_{days} (number of stocks, NL, number of days)
→Optimize LSTM hyperparameters
→Set output of Class 1 stocks to 1
→ Set output of Class 0 stocks to 0
→Execute equation 2 to 7 (deep learning)

Figure 2: LSTM Technique

3.2. Deep RankNet

Deep RankNet has been designed to learn the relative performance of documents/queries in a pairwise manner. It was pioneered by Burges et al. (2005). In their work, they use a probabilistic cost function, which utilizes a pair of sample objects to instigate and learn how to rank documents/objects. The cost function effectively seeks to minimize the number of pairing instances needed to correctly order a set of items. Given the challenges of financial asset predictions, Li and Tan (2021) bring forward a Deep RankNet to rank the performance of financial assets and ultimately bundle high-performing stocks into trading portfolios. The model instituted in this work will implement the framework introduced by Li and Tan (2021). However, we integrate novel elements of subperiods analysis. To start, an excess return is computed for a given stock relative to the composite index (market index). Here we contest 100 stocks all traded in the TSE. The equations for the return rate of a given stock, Rt,s, return of the index Rt,I, and excess return, ER (Li and Tan, 2021) are presented below.

$$R_t^s = \frac{P_t^s}{P_{t-1}^s} - 1 \tag{8}$$

$$R_t^I = \frac{P_t^I}{P_{t-1}^I} - 1 \tag{9}$$

$$ER_{t,m}^S = R_{t,m}^s - R_{t,m}^I \tag{10}$$

R = Return made by a given stock (s□S); P = Closing Price a stock (or index i); t = trading day; ER= Excess return of stock s relative to index i, for holding period m. Figure 3 illustrates the Deep RankNet procedure used in this work. The procedure follows closely the work of Li and Tan (2021).

→Compute excess return for all N_{stocks}
→Rank Stocks' returns from highest to lowest
→Assign top 50% stocks a ranking label of *1_Class 1*. Assign lowest 50% stocks a ranking label of *0_Class 2*
→Feed the *ER's* into a neural deep network as input
→Input Matrix is $N_{stocks} \times N_{days}$ (number of stocks, number of days)
→→Each stock entry includes fixed number of days (N_{days}), past *ER's*
→Initiate Neural Network outputs
→→*Score 1* for *Class 1* data
→→*Score 2* for *Class 2* data
→→→Subtract *score 2* from *score 1* to produce *Dif*
→→→Input *Dif* into a Sigmoid function
→Apply loss cross entropy function
→Apply gradient descent
→Update Neural Network parameters
→Output neural network scoring model

Figure 3: Deep RankNet procedure

3.3. XGBOOST (XGB)

XGBoost (XGB) stands for eXtreme Gradient Boosting. It was introduced by Chen (2016). The method achieves high accuracy and exceptionally fast speed due to its low computational complexity. The idea behind the method is to combine a penalty term and a loss function term. In turn, the method aims at obtaining high accuracy solution by minimizing the penalty term while preventing over-fitting via the reduction of the model's variance (Chen and Guestrin, 2016). The XGB method uses K additive function to predict output:

$$\widehat{y_i} = \varphi(x_i) = \sum_{k=1}^{K} f_k(x_k)\,,\ f_k \epsilon F \tag{11}$$

Where F is the space of regression trees and fk corresponds to each tree structure q and leaf weight w. The space of the regression tree can be written as:

$$F = \left\{w_{q(X)}\right\} for\ q: \mathbb{R}^m \to T, w \in \mathbb{R}^T \tag{12}$$

Where T is the number of leaves in the tree for *m* features. The objective function in the ensemble tree model is:

$$L(\varphi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{13}$$

Where l is a differentiable convex loss function and omega penalizes the complexity. Given the limitation of ensemble models as they cannot be optimized using traditional methods (in Euclidean space), the XGB is trained in an additive manner.

$$L(\varphi) = \sum_i l(\hat{y}_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{14}$$

Equation 14 integrates a greedy ft function, which improve the original model (eq. 13) significantly (Chen and Guestrin, 2016). This work optimizes the value of Gama, γ, in the $\Omega(f_t)$ term:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda\|w\|^2 \tag{15}$$

The work optimizes the values of γ, T (see eqn. 12), and F (eqn. 11 and 12). Figure 4 depicts the pseudo code of the XGB algorithm.

→Compute excess return for all Stocks

→Compute the median of each stock

→Assign top 50% stocks that are higher than the median to Class 1

→Assign lowest 50% stocks a ranking label of Class 0

→Construct Input Feature Matrix of (Nstocks*

Figure 4: XGBoost Procedure

3.4. Genetic Algorithm

Genetic Algorithm (GA) is an optimization and search technique based on the principles of genetics and natural selection. Pertaining to this work, the neural network's hyperparameters are numerous and quite difficult to tune. We start by populating an N random instances matrix for M-vector elements. We shall label each row as a chromosome, where each chromosome consists of a set of values for the hyperparameters in the question. We run the neural network for each chromosome and produce a unique prediction error. Then, we sort the chromosomes from lowest error value to highest. Next, the top half of our population table/matrix is chosen for mating (i.e., natural selection). Mating produces offspring that will populate the top of the table while the parents (top half of the previous iterate) will populate the bottom half. This way, if the parents turn out to be more effective than their offspring, they are not lost in the process (Goldberg, 1989; Holland, 1975). Then comes mutation where in each generation, we introduce 4% new random chromosomes to assure that a broader exploration of the solution space is warranted.

4. Results

The following subsections aim to provide a comprehensive analysis of the performance of the selected machine learning algorithms. This includes exploring the impact of hyperparameter

optimization on the algorithms' results, and evaluating the performance of different machine learning techniques.

4.1. Hyper-parameters

To demonstrate the impact of combining learning with optimization, we first present a random grid search. Our initial focus is solely on stocks traded on the Toronto Stock Exchange (TSE), for which we use 100 TSE stocks (closing prices obtained via Yahoo Finance's Python API) during the study period. The study requires complete data sets throughout all periods, thus several preprocessing steps were taken to eliminate financial assets with missing data, unresponsive updates (due to the API used), and inconsistencies. The hyper-parameters, to be optimized in this study, include the training ratio, batch size, number of neurons, stopping criterion, number of lags, number of hidden layers, and number of days in a time period. The batch size is a key factor that balances the speed of training with the quality of training. The number of neurons can add complexity to the model but may also result in overfitting. The stopping criterion sets the maximum number of non-improvements allowed in the solution before termination to avoid missing the global minimum. The number of hidden layers and lags also impact the predictions. The size of the time period in the training data can also have an impact on the results, with longer periods favored in the literature but shorter periods giving more emphasis on recent inputs. Figure 5 illustrates the correlation between these factors and DFs, which is the prediction error.

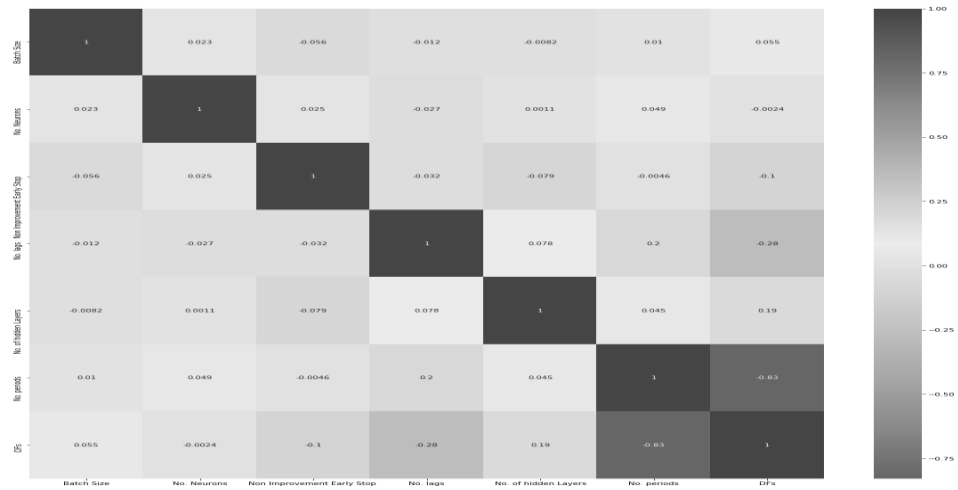


Figure 5: heat map of the correlations of the hyperparameter (FAP, DFs is the prediction error)

We utilize the genetic algorithm (GA) introduced in section 3 to optimize the hyperparameters (see figure 5). The training ratio of 80 to 20: 80% training and 20% turn to perform better and

is going to be used for the algorithms. Overall, the genetic-based algorithm does improve prediction accuracy. The hyperparameter optimization brings optimal value for a batch size of six, number of neurons of 22, and one hidden layer. For the XGB (XGBoost), the optimal value of gamma (eqn. 15) is 0.02. While the optimal learning rate occurs at 0.01. The optimal value for T (eqn. 11) is 300 and the optimal value of F (eqn. 11 & 12)is 8. Overall, we see 40% improvement in results with the use of the genetic technique.

4.2. Learning Algorithms

In this work, we bring add a novel element of feature engineering by aggregating data through the use of lags. We have found that this is especially beneficial for deep learning, as larger input data improves the training process. We employ period lags to manipulate the input, resulting in a higher quantity and quality of input data. Our preliminary research indicates that this feature engineering approach improves the accuracy of both XGB and LSTM algorithms.

The focus of this study is on portfolios with k=5 stocks. Three non-overlapping periods from 2019-01-01 to 2021-08-09 were analyzed. In the first two periods (as shown in Table 1), both XGB and LSTM outperformed the market in terms of mean return. Each period consisted of 200 days, approximately one calendar year. Notably, both techniques also outperformed the market during a negative revenue period in the second period. Despite the high variability in performance between the three techniques due to market fluctuations, all techniques performed better than the market during the initial wave of the COVID-19 pandemic, which is in line with previous studies (Bogomolov, 2013; Do and Faff, 2010; Huck, 2010) that have shown the effectiveness of pair trading strategies in periods of high turmoil. Looking across all three periods, we can see an apparent advantage of LSTM over the two other algorithms with an average daily return of 0.47% (compared to 0.29% for XGB). Alternatively, the Deep Rank performs poorly, with a negative daily return average. If we incorporate a 0.16% transaction cost as recommended in the work of Li and Tan (2021), we still see a marginal profit of 0.31% for the LSTM. This compares well with the work of Krauss et al. (2017). According to their analysis, they explain the very disappointing result from 2010 to 2015 as caused by an increase in public availability of powerful machine learning algorithms. If we assume the trend persists, the results, presented in the work, are quite significant since they still show overall positive daily return averages that are higher than the market.

Table 1: Three recent period comparisons between algorithms (200 days period)

	XGB	**LSTM**	**Deep Rank**	**Market**
Period from 2019-01-17 to 2020-01-06				
Mean	0.003066	0.003994	0.000377	0.000605
Standard Deviation	0.023545	0.033864	0.045080	0.006544
Sharpe ratio	1.455803	1.318577	0.093580	1.033376
Sortino ratio	2.559949	3.219885	0.208390	1.438714
Period from 2019-11-04 to 2020-10-21				
Mean	0.003454	0.002814	-0.004810	-0.000049
Standard Deviation	0.050973	0.054202	0.048303	0.027516
Sharpe ratio	0.757664	0.580408	-1.113287	-0.019919
Sortino ratio	1.365378	0.902510	-1.939949	-0.021003
Period from 2020-08-20 to 2021-08-09				
Mean	0.002358	0.007413	0.003212	0.002232
Standard Deviation	0.029439	0.051946	0.046531	0.009321
Sharpe ratio	0.895459	1.595566	0.771838	2.677102
Sortino ratio	1.760057	3.729138	2.044510	3.867835

Looking at the historic tail risk of the portfolios, based on the historic 1% and 5% VAR (value at risk), we measure the extent of potential extreme financial losses. These figures are not reported due to space limitation but can be provided upon request. Effectively, for all three periods (average), the 1%-VAR is higher for the LSTM standing at 9.7% while it is 8.0% for XGB illustrating a slight increase of risk while the 5%-VAR is quite similar between the two. The Sharpe ratio, which expresses the excess return per unit of risk, quantified in standard deviations, for LSTM is 0.13 points higher than for XGB. The Sortino ratio, which scales the returns by their downside deviation (Krauss et al., 2017) shows an upper hand for the LSTM with a ratio of 2.617 compared to 1.895 for the XGB.

Looking at figure 6, we can see the overall performance of the LSTM for three consecutive years, each year consisting of roughly 250 trading days. The Y-axis represents the daily return (not percentage), while the X-axis notes the period. In the figure, we observe a higher overall return in the LSTM model than the XGB. This correlates with the findings in Table 1, however, LSTM does exhibit higher variability than XGB. This is in line to the work of Chen et al. (2021).

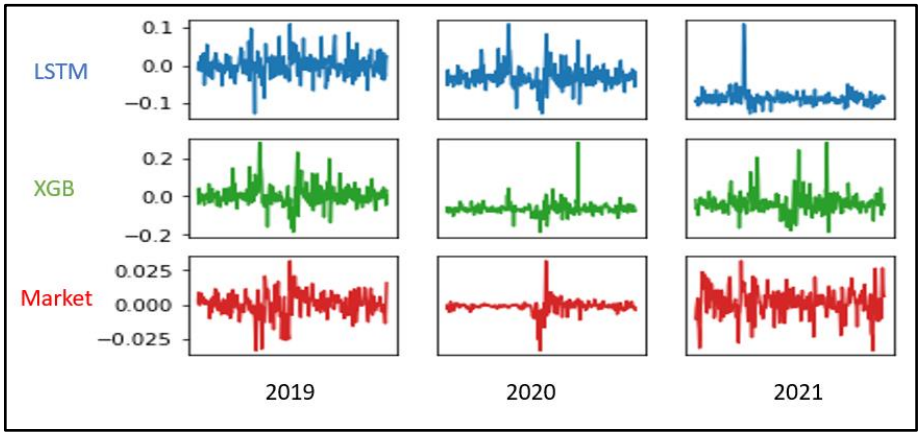


Figure 6: Average daily returns corresponding to year for XGB and LSTM

5. Conclusion and Recommendations

In this work, we utilize deep learning techniques to analyze financial data and make predictions for stock prices. Additionally, we employ ensemble decision trees to further enhance the accuracy of our predictions. The combination of these models is optimized through parameter tuning to achieve the best results for portfolio selection in trading. The ultimate goal is to optimize the holding period for maximum return on investment. The work studies a wide range of hyperparameters that affect the performance of the algorithms. We've taken the time to explicitly discuss many of the parameters in this work: Training ratio, Batchsize, Stopping criterion, period size, and others. The results indicate significant gains reaped by the optimizing the hyperparameters where the genetic algorithm brings 40% improvement compare to traditional random-grid approaches. To the best of our knowledge, this is the first work to coalesce Deep RankNet, LSTM, and a genetic-based algorithm.

Future research could focus on three key areas to further advance our understanding of algorithmic portfolio construction. Firstly, exploring different ranking rules for selecting stocks into portfolios would provide insights into the impact of different methods on portfolio performance. Secondly, more explanatory work is needed to understand why certain algorithms perform better than others in certain situations. This could involve detailed analysis of the underlying data patterns and market dynamics that drive algorithm performance. Lastly, expanding the scope of the research beyond North America to other financial markets would provide a more comprehensive understanding of algorithmic portfolio construction. This could involve testing the same algorithms on data from different financial markets and evaluating the performance in terms of return and risk characteristics.

Reference

Akita, R., Yoshihara, A., Matsubara, T., & Uehara, K.: Deep learning for stock prediction using numerical and textual information. IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), 1-6 (2016).

Basak, S., Kar, S., Saha, S., Khaidem, L., & Dey, S. R.: Predicting the direction of stock market prices using tree-based classifiers. The North American Journal of Economics and Finance, 47, 552-567 (2019).

Bogomolov, T.: Pairs trading based on statistical variability of the spread process. Quantitative Finance, 13(9), 1411–1430 (2013).

Burges, C., S. Tal., R. Erin., L. Ari., D. Matt., H. Nicole, and H. Greg.: Learning to Rank Using Gradient Descent. Proceedings of the 22nd International Conference on Machine Learning, 89–96 (2005).

Chen, T., & Guestrin, C.: Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm SIGKDD international conference on knowledge discovery and data mining, 785-794 (2016).

Chen, W., Zhang, H., Mehlawat, M. K., & Jia, L.: Mean–variance portfolio optimization using machine learning-based stock price prediction. Applied Soft Computing, 100, 106943 (2021).

Coqueret, G.: Persistence in factor-based supervised learning models. The Journal of Finance and Data Science, 8, 12-34 (2022).

Dessain, J.: Machine learning models predicting returns: Why most popular performance metrics are misleading and proposal for an efficient metric. Expert Systems with Applications, 199, 116970 (2022).

Do, B. , & Faff, R.: Does simple pairs trading still work? Financial Analysts Journal, 66(4), 83–95 (2010).

Du, J.: Mean–variance portfolio optimization with deep learning based-forecasts for cointegrated stocks. Expert Systems with Applications, 201, 117005 (2022).

Fischer, T., & Krauss, C.: Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 270(2), 654-669 (2018).

Goldberg, D.: Genetic algorithm in search, optimization and machine learning. Addison-Wesley, Reading, MA (1989).

Holland, J. H.: Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor (1975).

Huck, N.: Pairs selection and outranking: An application to the S&P 100 index. European Journal of Operational Research, 196(2), 819–825 (2009).

Huck, N.: Pairs trading and outranking: The multi-step-ahead forecasting case. European Journal of Operational Research, 207(3), 1702–1716 (2010).

Huck, N.: Pairs trading: Does volatility timing matter? Applied Economics, 47(57), 6239–6256 (2015).

Ismail, M. S., Noorani, M. S. M., Ismail, M., Razak, F. A., & Alias, M. A.: Predicting next day direction of stock price movement using machine learning methods with persistent

homology: Evidence from Kuala Lumpur Stock Exchange. Applied Soft Computing, 106422 (2020).

Jensen, M.: Some anomalous evidence regarding market efficiency. J. Financ. Econ., 6(2–3), 95-101 (1978).

Kim, S., Ku, S., Chang, W., & Song, J. W.: Predicting the Direction of US Stock Prices Using Effective Transfer Entropy and Machine Learning Techniques. IEEE Access, 8, 111660-111682 (2020).

Krauss, C., Do, X. A., & Huck, N.: Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. European Journal of Operational Research, 259(2), 689-702 (2017).

Kumar, I., Dogra, K., Utreja, C., & Yadav, P.: A comparative study of supervised machine learning algorithms for stock market trend prediction. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 1003-1007 (2018).

Li, Y., & Tan, Z.: Stock Portfolio Selection with Deep RankNet. The Journal of Financial Data Science, 3(3), 108-120 (2021).

Malkiel, B. G.: A Random Walk Down Wall Street. Norton, New York (1973).

Nelson, D. M., Pereira, A. C., & de Oliveira, R. A.: Stock market's price movement prediction with LSTM neural networks. In 2017 International joint conference on neural networks (IJCNN), 1419-1426 (2017).

Patel, J., Shah, S., Thakkar, P., & Kotecha, K.: Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert systems with applications, 42(1), 259-268 (2015).

Paiva, F. D., Cardoso, R. T. N., Hanaoka, G. P., & Duarte, W. M.: Decision-making for financial trading: A fusion approach of machine learning and portfolio selection. Expert Systems with Applications, 115, 635-655 (2019).

Porshnev, A., Redkin, I., & Shevchenko, A.: Machine learning in the prediction of stock market indicators based on historical data and data from twitter sentiment analysis. In 2013 IEEE 13th International Conference on Data Mining Workshops (pp. 440-444) (2013).

Reddy, V. K. S.: Stock market prediction using machine learning. International Research Journal of Engineering and Technology, 5(10) (2018).

Ren, R., Wu, D. D., & Liu, T.: Forecasting stock market movement direction using sentiment analysis and support vector machine. IEEE Systems Journal, 13(1), 760-770 (2018).

Shah, V. H.: Machine learning techniques for stock prediction. Foundations of Machine Learning, Spring, 1(1), 6-12 (2007).

Shen, S., Jiang, H., & Zhang, T.: Stock market forecasting using machine learning algorithms. Department of Electrical Engineering, Stanford University, Stanford, CA, 1-5 (2012).

Singh, R., & Srivastava, S.: Stock prediction using deep learning. Multimedia Tools and Applications, 76(18), 18569-18584 (2017).

Sharma, M., & Shekhawat, H. S.: Portfolio optimization and return prediction by integrating modified deep belief network and recurrent neural network. Knowledge-Based Systems, 109024 (2022).

Shi, S., Li, J., Li, G., Pan, P., Chen, Q., & Sun, Q.: GPM: A graph convolutional network based reinforcement learning framework for portfolio management. Neurocomputing, 498, 14-27 (2022).

Tim, J. M.: Artificial Intelligence–A System Approach. Computer Science Series, Infinity Science Press, 498 (2008).

Usmani, M., Adil, S. H., Raza, K., & Ali, S. S. A.: Stock market prediction using machine learning techniques. In 2016 3rd international conference on computer and information sciences (ICCOINS), 322-327 (2016).

Vazirani, S., Sharma, A., & Sharma, P.: Analysis of various machine learning algorithm and hybrid model for stock market prediction using python. In 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), 203-207 (2020).

Vijh, M., Chandola, D., Tikkiwal, V. A., & Kumar, A.: Stock Closing Price Prediction using Machine Learning Techniques. Procedia Computer Science, 167, 599-606 (2020).

Wang, G., Gunasekaran, A., & Ngai, E. W.: Distribution network design with big data: model and analysis. Annals of Operations Research, 270(1-2), 539-551 (2018).

Wu, D., Wang, X., & Wu, S.: A Hybrid Framework Based on Extreme Learning Machine, Discrete Wavelet Transform, and Autoencoder with Feature Penalty for Stock Prediction. Expert Systems with Applications, 118006 (2022).

Yadav, A., Jha, C. K., & Sharan, A.: Optimizing LSTM for time series prediction in Indian stock market. Procedia Computer Science, 167, 2091-2100 (2020).

Zhao, D., Bai, L., Fang, Y., & Wang, S.: Multi-period portfolio selection with investor views based on scenario tree. Applied Mathematics and Computation, 418, 126813 (2022).

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17761

Prediction of SMEs Bankruptcy at the Industry Level with Balance Sheets and Website Indicators

Carlo Bottai [1], Lisa Crosato[2], Caterina Liberati[1]

[1]Department of Economics, Management and Statistics, Università di Milano-Bicocca, Italy, [2]Department of Ecoomics, Ca' Foscari Universty of Venice, Italy.

How to cite: Bottai, C.; Crosato, L.; Liberati, C.; 2024. Prediction of SMEs bankruptcy at the Industry level with balance sheets and website indicators. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17761

Abstract

This paper addresses the importance of industry-specific models for SMEs bankruptcy prediction, building on earlier research finding larger predictive accuracy and enhanced temporal stability. Using Italian data, we propose separate bankruptcy prediction models for a few industries based on balance sheet data and explore the predictive power of SMEs' website HTML code structure. Our findings suggest that website data can serve as a valid complementary source for bankruptcy prediction, with different performances across sectors. We observe a certain degree of sectoral heterogeneity in the importance of balance sheet indicators and website structure, calling for an industry-tailored approach in bankruptcy prediction models.

Keywords: *website data, HTML code, SMEs, supervised learning.*

1. Introduction

The importance of exploring industry-specific variations in bankruptcy models was emphasized in earlier studies by Altman (1973), Altman and Izan (1984) and Platt and Platt (1990, 1991). They advocated for the use of industry-normalized company ratios as primary indicators in early-warnings for bankruptcy, asserting that industry-relative ratios effectively control for industry differences, resulting also in more temporally stable models. More recent research identified differences in variables influencing financial distress based on the technological level of the industry (Madrid-Guijarro et al., 2011). Bragoli et al. (2022) further proved the benefits of incorporating industrial variables into bankruptcy models, particularly in forecasting performance within the manufacturing sector.

An alternative approach in the literature has been the development of prediction models tailored to specific industries (Rikkers and Thibeault 2011, Ciampi and Gordini, 2013). Models estimated on miscellaneous industries have been found to underperform when applied on a

single industry (a case study for retail can be found in He and Kamath, 2006). Several works have focused on single sectors, aiming to uncover sector-specific factors influencing SME failure. This is based on the assumption that variables effective in one industry may not be applicable to others. For example, the construction sector is unique due to the extended duration of construction projects (Wang et al., 2024). Compared to other manufacturing industries, the food sector appears more susceptible to productivity shocks and less impacted by bank credit on default risk (Aleksanyan and Huiban, 2016).

Some studies have highlighted limitations of financial ratio variables, noting that they may reflect a company's recorded book value rather than its true value. Moreover, financial ratios may not encompass all information related to financial distress. Typically, financial ratios perform better in the manufacturing sector compared to retail, hospitality, and construction sectors. Concerns have also been raised about the potential impact of deliberate managerial actions distorting the financial situation (Serrano-Cinca et al. 2014, da Silva Mattos and Shasha, 2024). Consequently, researchers have started integrating other variables to enhance the predictive power of models, including non-financial characteristics and economic conditions of companies and industries. Some studies have explored the incorporation of website data, albeit on a limited sample basis (Blázquez et al., 2018, Crosato et al., 2021, Crosato et al., 2023).

Our paper contributes to this literature in two ways. Firstly, we present separate models predicting bankruptcy in six sectors at the 1-digit level of the NACE classification using balance sheet data from Aida, the Bureau Van Dijk (BvD) database describing Italian companies. Secondly, we investigate whether the HTML code structure of Small and Medium Enterprises' (SMEs) websites aids in predicting bankruptcy within the same sectors. The rationale behind utilizing website data is that financially sound firms would likely continue updating and maintaining their websites, while distressed firms might cease or neglect such activities.

Our results indicate that website data serve as a valid complementary source for bankruptcy prediction, with varying performances across sectors. We observe a certain degree of variability in the correct classification also when using balance-sheet data, although to a less extent. The variables selected by the stepwise algorithm also change, indicating sectoral heterogeneity in the importance of balance sheet indicators.

2. Data description

To proceed in our investigation, we combine information about each Italian SME (including any small business with NACE codes from C to N, except K, and incorporated by 2017) from two sources.

The data about each firm's characteristics (number of employees, industry, geographical location and website URL) and balance sheets are from Aida by BvD and refer to the year 2018,

for a total of 853,124 Italian SMEs. We define as defaulted by 2019 ($y = 1$) the 23,872 SME whose balance sheet was available up to 2018 and not in 2019; i.e., 2.8 per cent of the sample. Instead, we consider as survived to 2019 ($y = 0$) any SME lastly observed later than 2018 or that is still 'active' with no ongoing default procedure, according to BvD (829,252 SMEs).

Data about each firm's website are from the Wayback Machine by the Internet Archive (IA). For each SME, we pick the URL of its corporate website, if present on Aida; we download the snapshot closest to mid-June 2018, if archived by IA; and preserve as correct websites only those on which we detect the VAT number, phone number, postal code, or full address of the corresponding enterprise (see Blázquez et al., 2018 and Bottai et al., 2022 about this methodology).

To work on SMEs with both balance sheets and website information, our sample was reduced to 152,559 SMEs, of which 1.3 per cent defaulted by 2019. As shown in Fig. 1–3, firm size seems making some difference in terms of default probability. On the contrary, there is little evidence of industry- or location-specific effects, apart from exceptions like the manufacturing sector.

For each downloaded website, we build several size variables, capturing different aspects of the complexity of the website's homepage. Moreover, we extract the HTML tags from its code and count the number of times each website is used on the web-page.

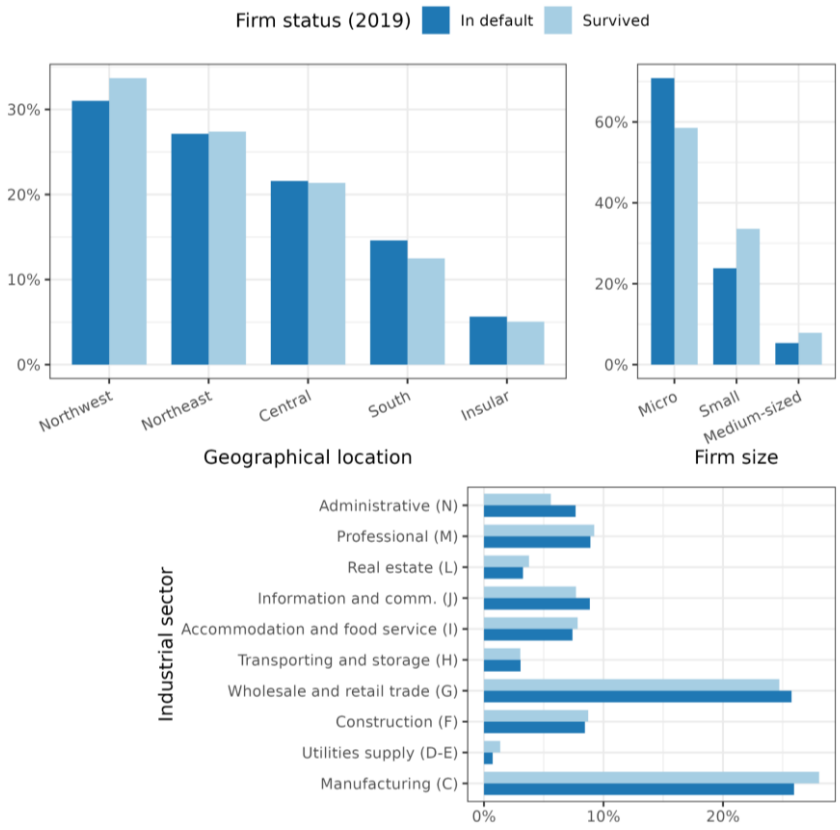


Figure 1. Percentage distributions of defaulted and survived SMEs by location, size and industry. Values computed on the 152,559 instances composing the final sample.

3. Results

We selected the NACE sections that had at least 1% of bankruptcies in 2019. These sectors include Manufacturing (C), Construction (F), Wholesale and retail trade; repair of motor vehicles and motorcycles (G), Transportation and storage (H), Information and communication (J), and Administrative and support service activities (N). Our dependent variable is default in 2019, whereas all predictors refer to the year 2018. For each sector, we divided the companies into two groups: training set (70%) and test set (30%). The training set was balanced to match the number of survived companies with the number of bankrupt ones, while the testing set remained unbalanced in favor of survived companies. Subsequently, we estimated stepwise logistic regressions separately on balance sheet and HTML indicators. We chose logistic regression because it is a well-known method that offers a straightforward interpretability.

Classification performances vary significantly from one sector to another, using either set of proposed variables (Table 1). Specifically, the overall classification measured by the geometric mean between sensitivity and specificity is very good for sectors C and G using both balance sheets and HTML code indicators. It is still fair in sectors H, J, and N when using balance sheet indicators and low, but still larger than 50%, in sectors F, J, and N when using HTML variables.

Table 1. Classification performances based on balance sheets or HTML indicators by industry (geometric mean of sensitivity and specificity).

Indicators	**C**	**F**	**G**	**H**	**J**	**N**
Balance sheets	0.783	0.783	0.759	0.731	0.683	0.712
HTML	0.780	0.540	0.750	0.591	0.538	0.532

Considering that HTML variables are available in real-time and completely free and accessible, the results in sectors C and G are noticeable. A probable reason behind the lower metrics within the remaining sectors lies in how companies in those sectors utilize their websites. For example, SMEs in sectors F and H most likely make poor use of their websites, hence they do not invest in the necessary technology. In other words, the limited discriminant power of HTML variables might be due to a general low level of website quality. The same reasoning can be extended to sectors J and N where, on the contrary, a high level of website quality is to be expected due to the need for well-functioning websites.

The balance sheet variables included by the model in each sector are collected in Table 2, where it can be seen that no variables are selected in all sectors. Out of the total 27 stepwise-selected variables, one is present in 4 sectors (ROE), four are present in three sectors (asset turnover, solvency ratio, profit and sales), eight in two sectors, and the remaining fourteen in one sector only. Therefore, most variables prove useful for discrimination in only one sector or the other. Notice that the set of discriminant variables for the manufacturing sector is the largest one (10 variables), followed by the sets for sectors G and N (9), H (8), and F and J (5 variables only).

In conclusion, these results highlight the importance of applying sector-specific prediction models. While working within sectors may have the disadvantage of a small number of failed companies to work with, considering a global model with dummy variables identifying sectors has the drawback of not highlighting sector-specific variables. Thus, they do not provide guidance to financial intermediaries on the aspects to focus on when evaluating firms in a sector. Furthermore, in a global model, classification metrics are usually reported at the aggregate, and not sector-specific, level. Further research should investigate in a similar fashion the role of HTML features, as well as estimate more complex models possibly combining the two types of variables. Augmenting HTML code with textual analysis could also improve classification metrics.

Table 2. Significant balance sheet variables (5% level) selected by the stepwise logistic regressions in each industry

	Industry					
Variable	**C**	**F**	**G**	**H**	**J**	**N**
Inventories	X					
Asset turnover	X			X		X
Asset	X					
Net working capital	X				X	
Tangible fixed assets	X					
Personnel cost	X		X			
EBITDA/Sales	X			X		
ROS	X					
Added value per employee	X			X		
Solvency ratio	X	X	X			
Cost of production services		X				
Total debts			X			
Profit		X	X			X
Revenue from sales		X	X			X
ROE		X	X	X	X	
Shareholder funds			X	X		
Financial fixed assets			X			
Long-term payable due to banks (yes)			X		X	
Intangible fixed assets				X		
Current assets				X		X
Added value				X		X
Short-term debt over the total debt					X	
EBIT					X	
Current ratio						X
EBITDA						X
Raw consumable materials and goods for resale						X
Wages						X
n. of relevant variables	10	5	9	8	5	9

References

Aleksanyan, L., & Huiban, J. P. (2016). Economic and financial determinants of firm bankruptcy: evidence from the French food industry. *Review of Agricultural, Food and Environmental Studies, 97*, 89-108.

Altman, E., 1973, Predicting railroad bankruptcies in America, Bell Journal of Economics and Management Science, 184-211.

Altman, E. and H. Izan, 1984, Identifying corporate distress in Australia: An industry relative analysis, Working paper (New York University).

Blázquez, D., Domènech, J., & Debón, A. (2018). Do corporate websites' changes reflect firms' survival? Online Information Review 42(6), 956–970. doi:10.1108/OIR-11-2016-0321.

Bottai, C., Crosato, L., Domènech, J., Guerzoni, M., & Liberati, C. (2022). Unconventional data for policy: Using Big Data for detecting Italian innovative SMEs. In Proceedings of the 2022 ACM Conference on Information Technology for Social Good, 338–344. New York, NY: Association for Computing Machinery. doi:10.1145/3524458.3547246.

Bragoli, D., Ferretti, C., Ganugi, P., Marseguerra, G., Mezzogori, D., & Zammori, F. (2022). Machine-learning models for bankruptcy prediction: do industrial variables matter?. *Spatial Economic Analysis*, *17*(2), 156-177.

Ciampi, F., & Gordini, N. (2013). Small enterprise default prediction modeling through artificial neural networks: an empirical analysis of i talian small enterprises. *Journal of Small Business Management*, *51*(1), 23-45.

da Silva Mattos, E., & Shasha, D. (2024). Bankruptcy prediction with low-quality financial information. *Expert Systems with Applications*, *237*, 121418.

He, Y., & Kamath, R. (2006). Business failure prediction in retail industry: an empirical evaluation of generic bankruptcy prediction models. *Academy of Accounting and Financial Studies Journal*, *10*(2), 97.

Madrid-Guijarro, A., Garcia-Perez-de-Lema, D. and van Auken, H. (2011), An analysis of non-financial factors associated with financial distress, Entrepreneurship & Regional Development, Vol. 23 Nos 3-4, pp. 159-186.

Platt, H. D., & Platt, M. B. (1990). Development of a class of stable predictive variables: the case of bankruptcy prediction. *Journal of Business Finance & Accounting*, *17*(1), 31-51.

Platt, H. D., & Platt, M. B. (1991). A note on the use of industry-relative ratios in bankruptcy prediction. *Journal of Banking & Finance*, *15*(6), 1183-1194.

Serrano-Cinca, C., Fuertes-Callén, Y., Gutiérrez-Nieto, B., & Cuellar-Fernández, B. (2014). Path modelling to bankruptcy: causes and symptoms of the banking crisis. *Applied Economics*, *46*(31), 3798-3811.

Wang, J., Li, M., Skitmore, M., & Chen, J. (2024). Predicting Construction Company Insolvent Failure: A Scientometric Analysis and Qualitative Review of Research Trends. *Sustainability*, *16*(6), 2290.

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17831

Violence Index: a new data-driven proposal to conflict monitoring

Luca Macis, Marco Tagliapietra, Elena Siletti, Paola Pisano
Department of Economics and Statistics "Cognetti de Martiis", University of Turin, Italy

How to cite: Macis, L.; Pisano, P.; Siletti, E.; Tagliapietra, M. 2024. Violence Index: a new data-driven proposal to conflict monitoring. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17831

Abstract

In this work, we propose a Violence Index (VI), as a comprehensive indicator of violence related to wars, conflicts, and disorders across different countries worldwide. This index is defined by mashing up different data sources: big data represented by the temporal progression of ACLED (Armed Conflict Location and Event Data) variables and an ad hoc dataset defined by our team documenting wars, armed conflicts, civil wars and violent demonstrations since 2010. The purpose is to encapsulate the intensity and impact of such unrest events in a single measure, the VI, to give a simpler, up-to-date, and manageable tool to practitioners and policymakers, for both prevention and strategic planning to let them behave better in future tragic scenarios.

Keywords: *ACLED; war; indicator; merging data; web scraping*

1. Introduction

Conflict prediction and early warning systems play a crucial role by identifying potential risks and threats, and offering decision-makers timely information to formulate policies for conflict mitigation and prevention. Scholars identified two possible ways of doing conflict and unrest prediction: the identification of potential conflicts and crises heavily relies on individual diplomatic and political knowledge, intuition, and subjective judgment; or you can recognize the potential of current technology, data science, and adopt it in this field (Gleditsch, 2002). In literature, different data have been adopted for this kind of issue, and they can be summarized in "Social Datasets" (i.e. social network data as Twitter, Telegram, etc.), "Disaggregated Datasets" (i.e. ACLED, UCDP (Uppsala Conflict Data Program), GDELT(Global Database of Events, Language, and Tone)) or "Aggregated Dataset" (i.e. World Bank, V-DEM).

Some scholars have enlightened some issues related to the use of "Social Datasets" in this context. The effectiveness of social network—Twitter—data for organizing insurrections has diminished due to prohibitions on violent tweets and the tracking of users by authoritarian regimes (Junior et al., 2021); There are no notable applications in the field of conflict prediction using social network data (Telegram, Khaund et al., 2021). Considering diplomatic datasets –

both disaggregated and aggregated – as ACLED (Raleigh et al., 2023), UCDP (Sundberg et al., 2013; Davies et al., 2023) or GDELT (Leetaru et al., 2013), other issues could occur: they are defined using indirect information (newspapers, etc.) and not by direct sources; moreover, they are mostly disaggregated datasets, since the disaggregation form avoids the use before data manipulation to practitioners or policymakers.

To avoid these issues, in our proposal, we adopted different datasets, with different characteristics in origin, structure, and frequency (Iacus et al, 2020), to define a single indicator (VI), which allows prompt use, without any further analysis, by policymakers.

2. The Data

The VI is defined mashing up different sources of data. In the first step, we use an ad hoc dataset defined by our team collecting in different ways information about the wars and armed conflicts that have occurred from 2010. In the second step we use a huge dataset, that is largely adopted for studies in the wars or conflict context (Hegre et al., 2012; Halkia et al., 2020), defined as the temporal progression of ACLED variables across more than 40 wars and conflicts spanning from the year 2010 to the present. In the following sections a description of the data.

2.1. ACLED Data

The Armed Conflict Location and Event Data Project (ACLED) is a project finalized for data collection, analysis, and crisis mapping, that was created by Clionadh Raleigh in 2005. It is a remote organization, which allows its team to live and work in all countries and contexts, where they collect and analyze instability. The members work within ACLED's executive office, global programs, external engagements, fundraising, and development or operations departments. In this way, ACLED data are derived from a wide range of local, national, and international sources in over 75 languages. The team conducts analysis to describe, explore, and test conflict scenarios, and makes both data and analysis open for free use. Moreover, researchers worldwide collect information on the dates, actors, locations, fatalities, and types of all reported political violence and protest events around the world. All data is updated in real time and published weekly. Years of historical coverage vary across countries and regions.

As detailed in Table 1, ACLED data take into account different event types, they focus on tracking a range of violent and non-violent actions by or affecting political agents, including governments, rebels, militias, identity groups, political parties, external forces, rioters, protesters, and civilians.

The total events collected by ACLED since 1997 are more than two millions, indeed from 2010 are more than one and a half million. The structure of the data within this dataset is meticulously designed to capture and organize information essential for comprehensive analysis of conflict

dynamics. At its core, ACLED data revolves around detailed event descriptions, encompassing the date, time, and location of each recorded incident. This information is vital for understanding the temporal and spatial dimensions of conflicts. Moreover, ACLED provides in-depth insights into the nature of events, including the actors involved and the characteristics of each incident. This categorization enables researchers and analysts to discern patterns of conflict, identify key stakeholders, and assess the intensity and outcomes of various events. Central to the reliability of ACLED data is its rigorous verification process, which documents the sources of information for each event. This transparency enhances the credibility and trustworthiness of the data, essential for informed decision-making and academic research. Furthermore, ACLED's temporal and geospatial dimensions add depth to its analytical capabilities. By organizing data chronologically and georeferencing event locations, ACLED empowers researchers to conduct temporal and spatial analysis, identifying temporal trends and spatial hotspots in conflict activity.

Indeed, in our study, we have chosen to aggregate these data on a weekly basis, grouping them according to event types. This approach entails tabulating the occurrence of events within a specific week in a given country, leveraging the geospatial information provided by ACLED. Consequently, this process yields a dataset wherein each row corresponds to a particular week in a country, including the count of events of a specific type transpiring during that week within that country.

Simply, we defined a new panel dataset, defined by countries and weekly frequency. The time series consists of reporting the counts of sub-event types that occurred in each country and their respective fatalities. To finalize the creation of our dataset, we opted to exclude data exhibiting more than 94% zeros, reflecting a significant absence of observations. Following this criterion, we retained data from 175 countries for subsequent analysis.

Due to the nature of the original ACLED information, which could be downloaded for free by each user through their API[1], and the dimension of the dataset, it could be defined as a Big Data resource.

2.2. Wars dataset

For the creation of the ad hoc war dataset, we adopted a web scraping technique, downloading data from Wikipedia. In this way, we obtained a .csv file containing information about wars, conflicts, and violent protests from January 1st, 2010 to December 31st, 2022 (from now on, unrest events).

1 Armed Conflict Location and Event Data Project (ACLED); https://acleddata.com

Table 1. ACLED Event Types (https://acleddata.com/)

Event type	Su-event type	Disorder type
Battles	Government regains territory	Political violence
	Non-state actor overtakes territory	
	Armed clash	
Protests	Excessive force against protesters	Political violence; Demonstrations
	Protest with intervention	Demonstrations
	Peaceful protest	
Riots	Violent demonstration	
	Mob violence	Political violence
Explosions/ Remote violence	Chemical weapon	
	Air/drone strike	
	Suicide bomb	
	Shelling/artillery/missile attack	
	Remote explosive/landmine/IED	
	Grenade	
Violence against civilians	Sexual violence	
	Attack	
	Abduction/forced disappearance	
Strategic developments	Agreement	Strategic developments
	Arrests	
	Change to group/activity	
	Disrupted weapons use	
	Headquarters or base established	
	Looting/property destruction	
	Non-violent transfer of territory	
	Other	

We considered this time window since the ACLED data collection from 2010 takes into account not only African countries.

The necessity to use this data is because ACLED data only records pure factual events but does not record political statements such as declarations of wars, revolutions, etc..

For each unrest event, we kept information about:

- Country: the country where the war happened.
- ISO 3166 code: the 3 letters unique code associated with each country.
- Starting date: the starting date of the unrest event.
- Ending date: the ending date of the unrest event. If the war is still ongoing, we use "present".

- War name: a label that recognizes the event.
- Type: a classification describing the event (Violent Demonstration, Armed Conflict, Civil War)
- Precise location: if available, it reports the specific area, such as a city, region, or country.
- War description: a brief note portraying the event and its actors and reasons.
- Link: the Wikipedia link reporting the event.

After the web scraping, the dataset was defined by 69 rows – i.e., events – but after a first human analysis, some events were dropped because there were no recorded battles – by ACLED – in the period under consideration (one week before and one week after the starting date). The final dataset consists of 46 unrest events from 2010 to 2022.

3. Methods and results

The VI integrates all ACLED variables, their weights determined through our analysis of the data within the war ad hoc dataset. Our proposal, for this reason, is defined in different steps.

Initially, we normalized all ACLED variables using the Min-Max method, re-scaling them to a range between 0 and 1 to ensure uniformity across variables (Mazziotta & Pareto, 2020). This normalization facilitated comparisons of variable fluctuations over time. Specifically, we examined the percentage change of variables within a two-week timeframe, spanning one week before to one week after the starting date of an unrest event. This process was defined by:

$$y_{i,j,k} = \frac{x_{i,j,k} - min_i(x_{i,j})}{max_i(x_{i,j}) - min_i(x_{i,j})} \quad (1)$$

where i represents the countries, j denotes the variables, and k indicates the weeks.

Subsequently, we analyzed the behavior of scaled variables within a two-week window surrounding each unrest event. By aggregating these variations across all events, we quantified their contributions and expressed them consistently as percentages.

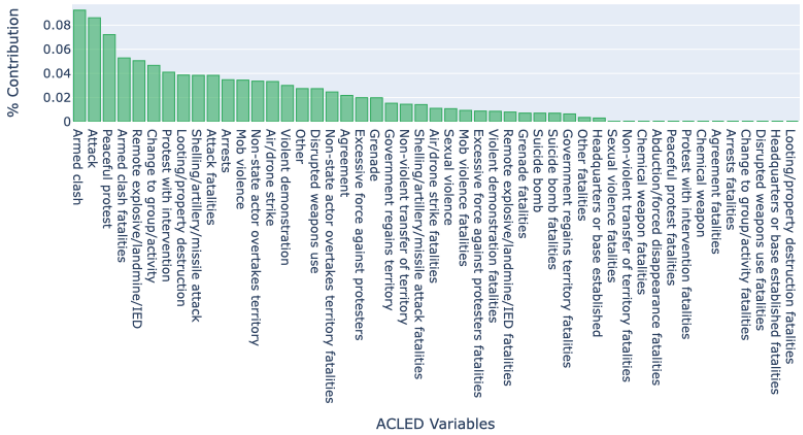


Figure 1. The contributions of each variable.

We then constructed the VI multiplying the scaled value of each ACLED variable by its corresponding weight contribution, determined in the previous step, for each week.

The culmination of our proposal involved summing the weekly contributions to obtain the VI. This index peaks during periods of war or armed conflict, reflecting intensified events among the 46 unrest events analyzed.

The summation of scaled data within the window allowed us to capture overall changes in scaled values during critical periods for each variable. Additionally, calculating the percentage contribution of each scaled variable within the specified time window provided insights into the variables with the most significant impact.

As a result, we can identify the variables that experienced the most substantial variations in intensity at the onset of wars since 2010. Lastly, we defined the VI for each country and day in the dataset based on scaled values of relevant variables and their respective percentage contributions. This calculation, outlined by equation (2), considers the weighted contributions of all ACLED variables, providing a quantitative description of unrest events tailored to each country.

$$VI_{i,j} = av_{i,j}^{(1)} \cdot cav^{(1)} + \quad av_{i,j}^{(2)} \cdot cav^{(2)} + \ldots + av_{i,j}^{(n)} \cdot cav^{(n)} \tag{2}$$

where:

$av^{(n)}$ is the *n*-th re-scaled ACLED variable, $cav^{(n)}$ is the *n*-th weighted contribution w.r.t. $av^{(n)}$, i is the *i*-th week, j is the *j*-th country, and n is the total number of ACLED variables.

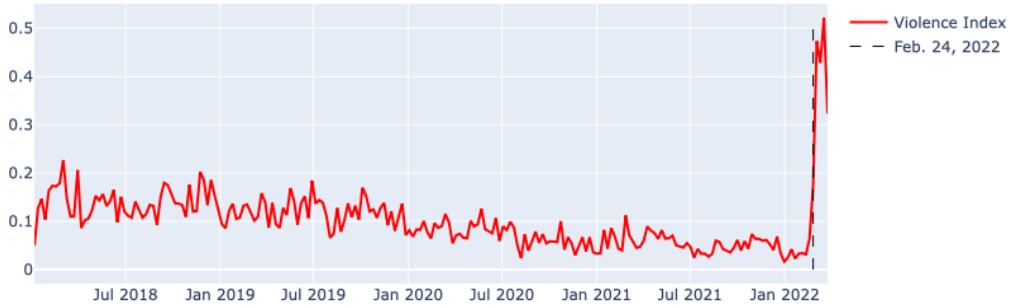


Figure 2.The VI trend from January 2017 to March 2022 in Ukraine.

The methodology offers several advantages. Firstly, it provides a detailed quantitative portrayal of unrest events specific to each country, thereby offering nuanced insights into the complexities of sociopolitical conflicts. Secondly, it allows for the delineation of unrest events tailored to the unique contexts of individual countries or regions. Furthermore, by integrating weighted contributions across all unrest events, the analysis ensures a comprehensive evaluation of variable impacts, enhancing the reliability of the results. However, the methodology also presents some challenges. The varying scales of magnitudes across countries may impede direct comparisons between countries, potentially complicating cross-country assessments. These considerations underscore the importance of cautious interpretation and contextualization of findings within the specific socio-political landscapes of each country.

Further results will be detailed and described during the presentation.

Acknowledgments:

This study was funded by the European Union - NextGenerationEU, in the framework of the GRINS -Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 – CUP D13C22002160001). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

References

ACLED. (2019). *ACLED Codebook, 2019*. Armed Conflict Location & Event Data Project (ACLED). www.acleddata.com

Halkia, M., Ferri, S., Schellens, M. K., Papazoglou, M., & Thomakos, D. (2020). The Global Conflict Risk Index: A quantitative tool for policy support on conflict prevention. *Progress in Disaster Science*, *6*, 100069. https://doi.org/10.1016/j.pdisas.2020.100069

Hegre, H., Karlsen, J., Nygård, H. M., Strand, H., & Urdal, H. (2012). Predicting Armed Conflict, 2010-20501. *International Studies Quarterly*, *57*(2), 250–270. https://doi.org/10.1111/isqu.12007

Davies, Shawn, et al. *Organized Violence 1989–2022, and the Return of Conflict between States.* Journal of Peace Research, vol. 60, no. 4, 13 July 2023, https://doi.org/10.1177/00223433231185169.

Gleditsch, K. S. (2002). Expanded Trade and GDP Data. Journal of Conflict Resolution, 46(5), 712–724. https://doi.org/10.1177/0022002702046005006

Iacus, S.M., Porro, G., Salini, S., & Siletti, E. (2020) Controlling for Selection Bias in Social Media Indicators through Official Statistics: a Proposal, *Journal of Official Statistics*, ISSN: 2001-7367 (0282-423X); doi: https://doi.org/10.2478/jos-2020-0017

Junior, M., Melo, P., Couto da Silva, A.P., Benevenuto, F. & Almeida, J. (2021). *Towards understanding the use of telegram by political groups in Brazil*. In: Proceedings of the Brazilian Symposium on Multimedia and the Web WebMedia '21 , 237–244. New York, NY, USA. Association for Computing Machinery (https://doi.org/10.1145/3470482.3479640).

Leetaru, K., & Schrodt, P. A. (2013, April). Gdelt: Global data on events, location, and tone, 1979–2012. In ISA annual convention (Vol. 2, No. 4, pp. 1-49). Citeseer.

Khaund, T., Hussain, M.N., Shaik, M. & Agarwal, N. (2021). Telegram: Data collection, opportunities and challenges. In: Lossio-Ventura, J.A., Valverde-Rebaza, J.C., Diaz, E. & Alatrista-Salas, H. (eds) *Information Management and Big Data* , 513–526. Cham. Springer International Publishing.

Mazziotta M., Pareto A. (2020) *Gli indici sintetici*, Giappichelli Ed., Turin.

Raleigh, C., Kishi, R. & Linke, A. (2023) Political instability patterns are obscured by conflict dataset scope conditions, sources, and coding choices. *Humanit Soc Sci Commun* 10, 74. https://doi.org/10.1057/s41599-023-01559-4

Sundberg, Ralph, and Erik Melander. *Introducing the UCDP Georeferenced Event Dataset.* Journal of Peace Research, vol. 50, no. 4, July 2013, pp. 523–532, ucdp.uu.se/downloads/ged/ged172.pdf, https://doi.org/10.1177/0022343313484347.

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17467

Potential of ChatGPT in predicting stock market trends based on Twitter Sentiment Analysis

Ummara Mumtaz, Summaya Mumtaz
Department of Information Technology, University of the Cumberlands, United States.

How to cite: Mumtaz, U.; Mumtaz, S.; 2024. Potential of ChatGPT in predicting stock market trends based on Twitter Sentiment Analysis. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17467

Abstract

The rise of ChatGPT has brought a notable shift to the AI sector, with its exceptional conversational skills and deep grasp of language. Recognizing its value across different areas, our study investigates ChatGPT's capacity to predict stock market movements using only social media tweets and sentiment analysis. We aim to see if ChatGPT can tap into the vast sentiment data on platforms like Twitter to offer insightful predictions about stock trends. We focus on determining if a tweet has a positive, negative, or neutral effect on two big tech giants Microsoft and Google's stock value. Our findings highlight a positive link between ChatGPT's evaluations and the following day's stock results for both tech companies. This research enriches our view on ChatGPT's adaptability and emphasizes the growing importance of AI in shaping financial market forecasts.

Keywords: *ChatGPT; Twitter; Sentiment Analysis; Stock Market Price; Trend Prediction; LLMs.*

1. Introduction

ChatGPT represents one of the latest advancements in the realm of artificial intelligence. It is grounded on the Generative Pre-trained Transformer (GPT) architecture (Vaswani et al., 2017), which enables it to understand and generate human-like text based on the input it receives. With its ability to engage in detailed, coherent, and contextually relevant conversations, ChatGPT has become a significant player in the AI industry (Gabashvili,2023). However, their role in financial economics, especially in predicting stock market returns, is still a new area of exploration. While some might argue that these general-purpose models, not being specifically designed for stock prediction, may not be particularly useful, others believe that their vast training on extensive text data and ability to grasp natural language context could make them valuable for this purpose. The actual ability of LLMs in forecasting financial trends remains uncertain. Our research aims to address this by analyzing how well ChatGPT can use sentiment analysis to predict stock market returns.

In the digital age, the traditional means of predicting stock market trends—like analyzing quarterly reports or market fundamentals—have been complemented with more novel approaches, among which sentiment analysis stands prominent. Sentiment analysis harnesses the vast amount of unstructured data on the internet to gauge public sentiment, thereby offering insights into potential market movements. A particularly vibrant source of this sentiment data is Twitter, where millions of users express their opinions on a plethora of subjects, including the stock market, every day. ChatGPT, a state-of-the-art language model, presents a significant leap in processing and understanding such unstructured data. Its capability to comprehend context and nuance in textual content makes it a prime candidate for analyzing Twitter sentiments related to stock market trends. But the challenge does not end at merely capturing sentiments; the ability to make predictions without being explicitly trained on specific tasks is a paramount advantage. This is where the zero-shot learning technique comes into play. Zero-shot learning (ZSL) allows models to make predictions or categorizations on data for which they have not seen any examples during training (Lampert et al.,2014). In the context of machine learning, ZSL is often used in situations where labeled data for some tasks is scarce or unavailable. The integration of zero-shot learning in chatbot technologies like ChatGPT demonstrates an advancement that allows these models to respond accurately to a wide range of user prompts, even if they've never encountered them during training. In the context of stock market prediction using ChatGPT, it means that even if the model has not been explicitly trained on stock market data, it can leverage its generalized understanding of language to assess sentiments and make predictions on stock trends. This approach is not only cost-effective but also incredibly versatile, accommodating rapid shifts in market dynamics or unforeseen events which might not be well-represented in training data.

In this study, we explore the potential of using ChatGPT in predicting stock market trends solely based on Twitter sentiment analysis and by employing the zero-shot learning strategy. By enriching the potent capabilities of ChatGPT with the richness of sentiment data on Twitter, we aim to chart a novel path in stock market trend prediction and provide insights into the potential and limitations of such an approach. The rest of the paper is organized as: section 2 gives an overview of the existing literature; Section 3 describes the methodology including the data collection, pre-processing and temporal prediction modeling technique. Section 4 includes the conclusion and future directions.

2. Literature Review

With the advent of social media, researchers have explored the possibility of utilizing the vast amount of user-generated content as a predictive tool for stock market movements. We aim to consolidate key findings from various studies that have investigated the potential of using social media tweets, specifically from platforms like Twitter, to forecast stock market trends. Bollen et al. (2011) were among the pioneers to explore the relationship between Twitter sentiment and

the stock market. They found a notable correlation between specific mood dimensions extracted from tweets and the Dow Jones Industrial Average. Zhang et al. (2011) also confirmed a significant relationship between Twitter sentiment and stock market movements but emphasized the importance of using sophisticated sentiment analysis tools. In recent years, advancements in machine learning have provided researchers with sophisticated tools. Nguyen et al. (2015) used Support Vector Machines (SVM) on Twitter data to predict stock prices and achieved a higher accuracy rate than traditional methods. Rao and Srivastava (2012) highlighted the importance of feature selection in sentiment analysis for accurate stock market prediction. While the sentiment of tweets is essential, the sheer volume of tweets mentioning a specific stock or related term can also be a predictor. Mao et al. (2012) found that the number of tweets about a company correlates positively with its trading volume. Siganos et al. (2014) argued for the advantages of real-time tweet analysis over daily aggregation, suggesting that intraday tweet volumes and sentiment shifts provide more immediate and actionable insights for traders.

We found two relevant studies that recently tried to explore the potential of ChatGPT in stock market trend prediction. The research conducted by (Lopez-Lira et al., 2023) explored the capabilities of ChatGPT and other large language models in forecasting stock market returns based on news headlines. By classifying headlines as positive, negative, or neutral for stock prices using ChatGPT, a significant positive link between the model's scores and subsequent daily returns was identified. While ChatGPT surpasses conventional sentiment analysis techniques, basic models like GPT-1, GPT-2, and BERT lack precision in predicting returns. Notably, strategies leveraging ChatGPT-4 yield the highest Sharpe ratio. The study also reveals consistent underreaction in the market to company news, with stronger predictability in smaller stocks and those with negative news, hinting at the role of limits-to-arbitrage. The study by Xie and colleagues (Xie et al., 2023) found that ChatGPT, falls short in this financial context, lagging behind both cutting-edge and traditional forecasting methods, like linear regression. Even with advanced prompting strategies and incorporating tweets, the model's performance is lacking, further revealing issues in its explain ability and consistency. These findings underscore the potential need for model fine-tuning and pave the way for future research that combines social media sentiment with stock data to refine financial market predictions. Not all studies found a consistent, strong correlation between Twitter sentiment and stock movements. Some, like Luss and d'Aspremont (2015), warned of potential overfitting when relying heavily on social media data, advocating for a mixed-methods approach. There's also the challenge of 'noise' in social media data. With the prevalence of bots and irrelevant content, filtering out noise remains a critical step in the analysis (Chen et al., 2018). The interplay between social media sentiment and stock market movements also raised questions about market efficiency and the potential for manipulation, as explored by researchers like Cook et al. (2018). Instead, our focus in this study is to evaluate whether ChatGPT, not trained in predicting returns, has the potential to predict stock market returns based only on tweets without considerable effort of

cleaning the tweets and training or finetuning the model. Through a simple approach that leverages the model's stock market trend prediction capabilities, using sentiment analysis on tweets data and compare it to the actual stock market trends.

3. Methodology

3.1. Data Collection & Pre-processing :

We utilized a publicly available dataset at Kaggle named "500k ChatGPT-related Tweets Jan-Mar 2023"[1] for this analysis. The original dataset is composed of 500K tweets from January 4th, 2023, to March 29th, 2023, which were extracted by searching term "gpt" in the tweets. Tweets were specifically sought that mentioned or related to three distinct terms: 'ChatGPT', 'Microsoft', and 'Google'. This targeted approach ensured the relevance of the tweets to the topic of interest. We performed minimal data cleaning and only removed URL present in the tweets. Apart from URL removal, no other noise or extraneous data such as emotion icons or hashtags was filtered out, preserving the authentic voice and sentiment of the tweets.

Stock performance data for Google and Microsoft was procured from the NASDAQ website. As with the tweets, the data captured was between January 4th, 2023 and February 28th, 2023, ensuring synchronicity in the dataset's temporal scope. The stock market percentage change indicates the relative change in the value or price of a stock, index, or any other relevant market indicator over a specific period. It provides a standardized way to understand how much a stock, or the overall market, has increased or decreased, allowing for easy comparison over time or against other assets.

From the assembled tweet collection, samples were structured not based on the number of tweets but on the number of tokens. For each date, we picked a random sample consisting of a collection of tweets that cumulatively contained approximately 15,000 tokens. This method was chosen to standardize the volume of textual content processed by ChatGPT for each date and to ensure ChatGPT allowed token limit does not exceed.

3.2. Sentiment Analysis and Stock Market Prediction using ChatGPT

Instruction-Based Prompt Engineering Input: For sentiment analysis and trend prediction, ChatGPT was provided with a single consolidated input for each date. The input format was initiated with an instruction, outlining the task, followed by the tweets for that date. For instance, one example instruction is: "Based on the following tweets, predict the stock market trends for

[1] https://www.kaggle.com/datasets/khalidryder777/500k-chatgpt-tweets-jan-mar-2023

Google and Microsoft. We also provide instructions without using explicit company names as shown in figure 1.

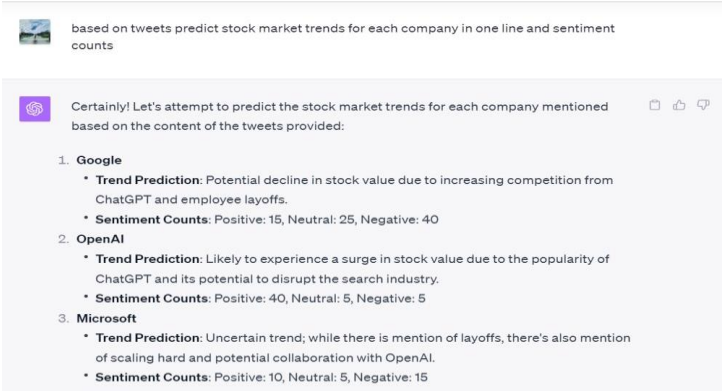


Figure 1 : Example prompt 1 for stock market trend prediction and ChatGPT results.

ChatGPT was able to identify all the companies in the given tweet along with sentiment counts for each company. For each date, we collected the ChatGPT predictions in an excel file for both companies, along with the associated sentiment counts. We observed that without any fine tuning or few-shot learning, ChatGPT was able to bring relevant list of companies and the associated tweets count based on sentiments.

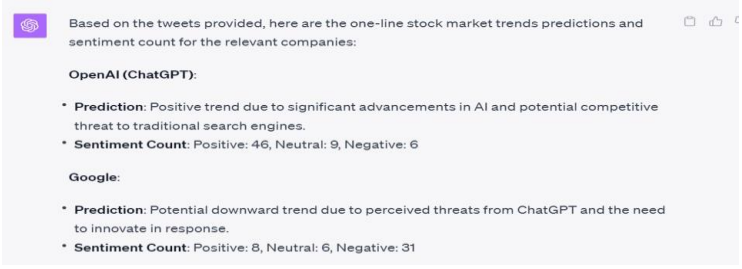


Figure 2 : Example prompt 2 for stock market trend prediction and ChatGPT Results

3.3. Temporal Predictive Modelling

We use temporal predictive modelling by utilizing a simple forward-looking approach. The stock market trend prediction from tweets on a specific day was utilized to predict stock market behavior for the subsequent day—before the market's opening. This was premised on the theory that today's public sentiment might influence tomorrow's stock performance. After obtaining ChatGPT's predictions for each date, these were systematically compared with the actual percentage changes in Google's and Microsoft's stock values as reported by NASDAQ. The objective was to ascertain the accuracy of the model's predictions and to determine the extent to which sentiment analysis could be used as a predictive tool for stock market trends. The model predicts the stock market trend in one to two sentences which cannot be directly compared with

the percentage change in shares. To compare the results, we used a human evaluation approach and manually matched the stock market trend predicted by ChatGPT, with the stock percentage change of both companies on the following dates.

3.4. Results

To provide a more intuitive understanding of the results, we plot the ChatGPT predictions along with the actual % Change on the same date to visualize the results. Figure 3. shows the predictions made for Microsoft. The x-axis shows the prediction made on a specific date and y-axis refers to actual % change on the same date. For Microsoft, out of a 37-day period, ChatGPT's predictions aligned with the actual outcomes on 26 occasions. Thus, getting an accuracy of 70% for Microsoft stock market trend prediction.

In assessing the predictive prowess of ChatGPT for Google's stock trends, it is found that the predictions deviated slightly from the actual outcomes. Specifically, out of a 36-day timeframe, the model made accurate forecasts on 23 days, resulting in an accuracy of roughly 63.88%. This accuracy, however, should be contextualized. Considering ChatGPT was not exclusively fine-tuned for stock market predictions and employing a zero-shot learning strategy, its performance is notably superior to a model making random predictions. The model showcased an ability beyond mere trend prediction; it attempted to identify the underlying reasons for certain trends. As illustrated in figure 4 concerning Google, ChatGPT not only made trend forecasts but also pinpointed specific factors like Google being "challenged by emerging competitors" or potential impacts from "Google’s bard", and even identified a "potential threat from ChatGPT" itself. This reflects ChatGPT's capability to not just foresee the market direction but also recognize pivotal elements influencing that direction. Furthermore, it's crucial to emphasize that the analysis was conducted using a restrained set of tweets, hinting that a more exhaustive data set from Twitter or other social platforms might yield even better insights.

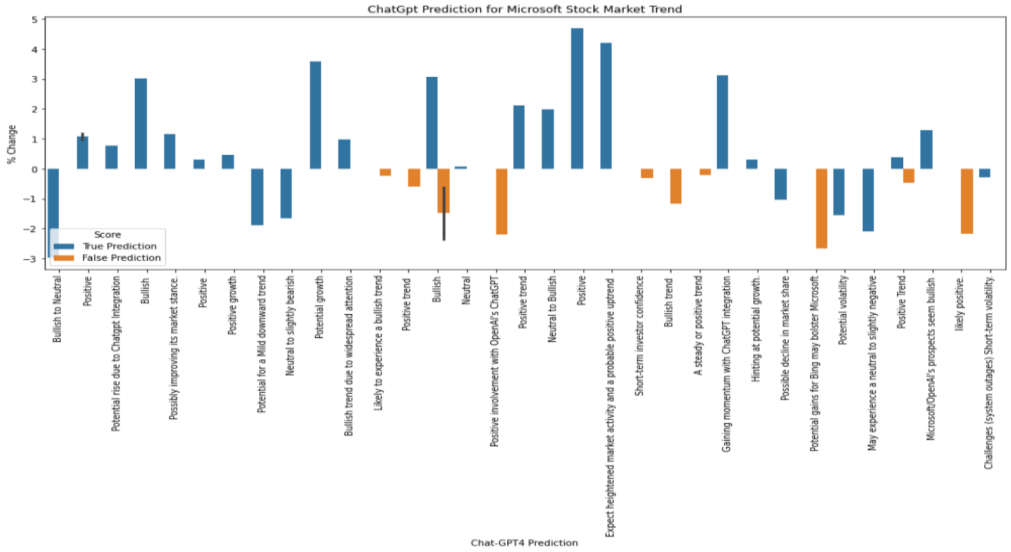


Figure 3: ChatGPT Result Comparison for Microsoft with actual % Change (Blue color signifies correct predictions and orange color shows incorrect predictions). Bars are organized from left to right starting from January 5th, 2023, to February 28th, 2023.

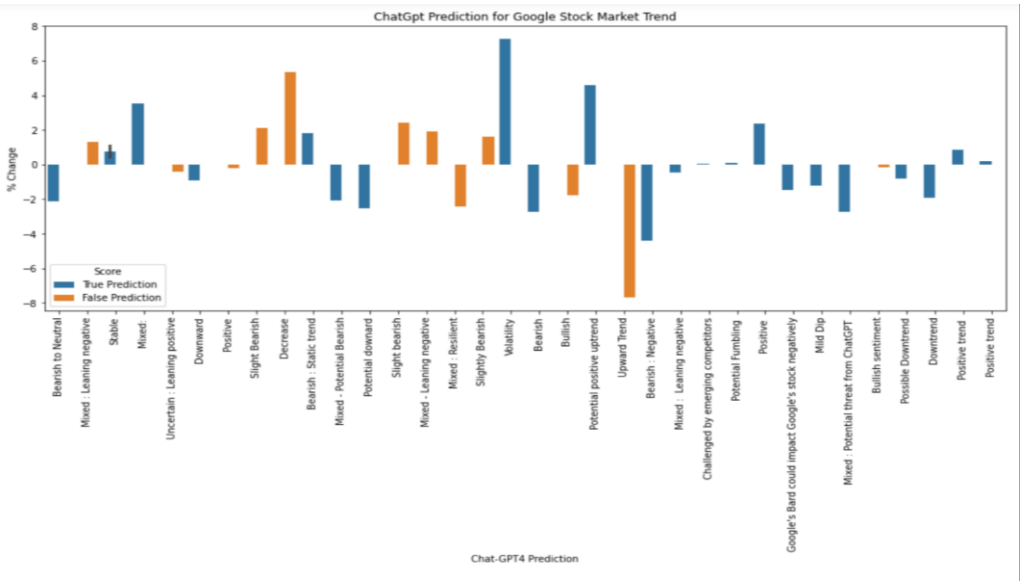


Figure 4 : ChatGPT Result Comparison for Google with actual % Change (Blue color signifies correct predictions and orange color shows incorrect predictions). Bars are organized from left to right starting from January 5th, 2023, to February 28th, 2023.

4. Conclusion and Future Work

We assessed ChatGPT's capability in predicting stock market trends for two major tech giants, Microsoft, and Google, utilizing only tweets and sentiment analysis. The results indicate that ChatGPT achieved an accuracy of 70% for Microsoft and 63.88% for Google. While the predictions were not always spot-on, it is crucial to highlight that ChatGPT was not primarily designed for stock market predictions. Yet, its performance markedly exceeded that of a randomly predicting model. It also demonstrated its ability to identify underlying factors influencing those trends, adding depth and context to its forecasts. Given ChatGPT's encouraging performance, future endeavors could revolve around leveraging a more comprehensive dataset, possibly tweets, and data from other social media platforms to enhance predictive accuracy. It would be intriguing to observe how a richer dataset affects the model's forecasting capacity. Additionally, insights into ChatGPT's ability to recognize and identify pivotal factors influencing market directions can be of immense value. It paves the way for more insightful and informed stock market predictions, which can be pivotal for traders.

References

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., and Polosukhin, I., (2017). Attention is all you need. NIPS 2017.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8.

C. H. Lampert, H. Nickisch, and S. Harmeling, (2014). Attribute-based classification for zero-shot visual object categorization, IEEE Trans. Pattern Anal. Mach. Intell., vol. 36.

Chen, H., De, P., Hu, Y. J., & Hwang, B. H. (2018). Wisdom of crowds: The value of stock opinions transmitted through social media. The Review of Financial Studies, 31(9).

Cook, S., Conrad, C., & Krauss, J. (2018). Underreaction to news in the US stock market. Quantitative Finance, 18(1), 45-56.

Gabashvili, I. S. (2023). The impact and applications of ChatGPT: a systematic review of literature reviews. arXiv preprint arXiv:2305.18086.

Lopez-Lira, A., & Tang, Y. (2023). Can chatgpt forecast stock price movements? return predictability and large language models. arXiv preprint arXiv:2304.07619.

Luss, R., & d'Aspremont, A. (2015). Predicting abnormal returns from news using text classification. Quantitative Finance, 15(6), 999-1012.

Mao, Y., Wei, W., Wang, B., & Liu, B. (2012). Correlating S&P 500 stocks with Twitter data. Proceedings of the 1st ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research. ACM.

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. Expert Systems with Applications, 42(24), 9603-9611.

Rao, T., & Srivastava, S. (2012). Using Twitter sentiment to forecast stock market movements. An Analytical Study.

Siganos, A., Vagenas-Nanos, E., & Verwijmeren, P. (2014). Facebook's daily sentiment and international stock markets. Journal of Economic Behavior & Organ., 107, 730-743.

Xie, Q., Han, W., Lai, Y., Peng, M., & Huang, J. (2023). The Wall Street Neophyte: A Zero-Shot Analysis of ChatGPT Over MultiModal Stock Movement Prediction Challenges. preprint arXiv:2304.05351.

Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting stock market indicators through Twitter "I hope it is not as bad as I fear". Procedia-Social and Behavioral Sciences, 26, 55-62.

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17158

Google Trends Forecasting of Youth Unemployment

Nathan de Bruijn, Fons Wijnhoven , Robin Effing
Behavioural, Management and Social Sciences, University of Twente, The Netherlands.

How to cite: De Bruijn N.; Wijnhoven F.; Effing R. 2024. Google Trends Forecasting of Youth Employment. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17158

Abstract

The forecasting field has been using the surge in big data and advanced computational capabilities. This article discusses the methodological issues of Google Trends (GT) data reliability and forecasting validity for youth unemployment forecasts. We demonstrate the problems with static GT forecasting procedures and show a 44% increase in forecasting accuracy by applying time-varying model respecification forecasting.

Keywords: *Forecasting; time series; rolling window; expanding window; unemployment; google trends*

1. Introduction

The complexity of using GT data in forecasting is reflected in the challenge of selecting suitable keywords from an array of millions of potential keywords (Varian, 2014). This complexity is amplified since GT data structurally changes over time (Behnen et al., 2020). This makes predicting with GT from both data reliability and prediction validity perspectives challenging. Therefore, our research question is "How can GT data reliably be used for generating valid unemployment predictions?" While answering this question, this study aims at contributing to insights on managing GT data reliability and forecasting in time-variant contexts.

In this article we first discuss literature related to benefits and drawbacks of using GT data. Subsequently, the methodology section outlines the research design. The results section presents our forecasts. Finally, we present our conclusions and discuss their implications.

2. Literature review

Choi & Varian (2012) demonstrated that the popularity of Google searches like "apply for unemployment" are useful in forecasting future unemployment. Similarly, Ginsberg et al. (2009) launched a tool called Google Flu Trends to forecast flue occurrences, but Google Flu Trends faced notable criticism when it overestimated doctor visits by a factor of more than two (Lazer et al., 2014). One of the causes was Google Flu Trends' continuous search for the most correlated keywords, without theorizing ex-ante which keywords (i.e., predictor variables) are

appropriate (Lazer et al., 2014). Moreover, algorithmic changes caused Google Flu Trends to become less accurate over time and GT was phased out by Google (Lazer et al., 2014).

The use of GT data does come with several advantages however. GT data is fully anonymized and data collection occurs without any effort from the user. Users may not even be aware of data collection, ensuring that the recorded data is unobtrusive and reflects natural behaviour (McLaren & Shanbhogue, 2011). Other advantages include the real-time availability of pre-processed data at no cost (Zhu et al., 2012). Nevertheless, pre-processing by Google does come with sampling error (Cebrián & Domenech, 2022). GT data is also criticized to be unreliable, due to the homogeneity of internet users having an influence on keyword popularity. Also, individuals may lack internet access or use alternative sources of search, resulting in GT coverage bias (Cebrián & Domenech, 2022).

Recent literature also discussed the inconsistency of GT data. For example, GT data for the same keyword, during the same period, may be different when again collected tomorrow (Cebrián & Domenech, 2022). Eichenauer et al. (2021) attribute this inconsistency to sampling variation, which is more noticeable for less popular keywords and smaller regions due to smaller samples. Furthermore, GT data has the tendency to structurally change over time (Behnen et al., 2020) giving rise to the issue of parameter instability in forecasting. Furthermore, spurious correlations between GT data and phenomena that need to be forecasted are found. For example, GT data for a popular drink was highly correlated to housing sales in the USA (Tran et al., 2017).

Even with these limitations, GT provides researchers with a large dataset of user behavior related to real world developments, useful if the data reliability and forecasting validity issues are handled well.

3. Methodology

3.1. Unemployment forecasting with GT

There are many studies that leveraged GT data to forecast unemployment focused on a single keyword. Examples being: D'Amuri & Marucci (2017) who used the keyword "jobs", Mclaren & Shanbogue (2011) who used "jobseeker's allowance", Simionescu & Cifuentes-Faura (2022) who used "unemployment", Naccarato et al. (2018) who used "job offers", Fondeur & Karamé (2013) who used "employment", and Vicente et al. (2015) who used "job offer". Other studies leveraged multiple keywords. For example, Tuhkuri (2016) created an index by averaging over thirteen keywords with weights based on a Google search volumes. However, the aforementioned studies relied on intuition, and no formal keyword selection techniques were used. Other studies adopted formal techniques for keyword selection, like Borup & Schutte's (2020) regularization approach and Singhania & Kundu (2021) who inputted over 500 potential keywords to a neural network. Also, Li et al. (2015) applied dimension reduction techniques to select keywords.

3.2. Data collection

The GT dataset we use for this study is the GT search volume index (SVI). A low SVI indicates a low search volume for the search keyword. Either weekly or monthly data can be obtained from GT. GT data spanning from April 2008 till December 2022 is collected for this study, which is in line with the time length of the unemployment dataset that we gained from the Dutch Census Office CBS. To deal with data invalidity, this study started from a domain ontology (Guizzardi et al., 2022), to then only select keywords based on literature and economic reasoning. This was carried out by the following steps.

1. To ensure that keywords are representative of the total number of unemployment related searches, we estimated the monthly search volume with the Google Ads Keyword Planner. Keywords with a low monthly search volume were considered not representative.
2. 26 studies that forecasted unemployment with GT data were studied to find the keywords they used. 329 keywords were obtained and checked for fitting into the domain ontology of unemployment. This led to the identification of 20 main themes, six of the prominently used are: (1) job search, (2) unemployment interest, (3) employment agency, (4) job platform, (5) unemployment benefits, and (6) unemployment claims. These six themes were used to find their Dutch equivalents. Additionally, Dutch keywords were taken from (Te Brake, 2017). Keywords with less than 1,000 monthly searches were removed. This resulted in 58 remaining keywords.
3. These 58 keywords were used to prompt Google's algorithm to return closely related keywords. The search volumes of the closely related keywords were also checked and after doing so another 21 keywords were added, resulting in a total of 79 keywords that accounted for 2,799,400 monthly Google searches.
4. The GT data for each keyword was obtained at 12 different moments across 9 days to reduce sampling variation. This resulted in 12 datasets for each keyword, summing to a total of 948 datasets.
5. The mean correlation between the 12 GT data samples for the same keyword was checked. All keywords with a mean correlation lower than .90 were dropped. Consequently, 63 keywords remained fit for analysis. For each of the 63 keywords that remained, the 12 datasets obtained for these keywords were averaged, as suggested by Eichenauer et al. (2021).

3.3. Analysis

Two procedures exist for out-of-sample forecasting; the fixed-origin and the rolling-origin (Hewamalage et al., 2023). While the fixed-origin procedure for out-of-sample forecasting has long been used, the rolling-origin procedure for out-of-sample forecasting has become the favorite choice. With the fixed-origin procedure, the complete dataset is divided into a training set and test set, and this division remains constant, i.e., all following predictions being based on the same training window. For the rolling window approach, the window size may be extended

(expanding window) or the window size may remain the same but each prediction of a following period may be on a later start of the training window (rolling window). For the actual predictions two approaches are commonly used (Hewamalage et al., 2023): 1) updating the forecasting model by feeding it with new data and thus without any change of the prediction model, 2) using new data to recalibrate the forecasting model, i.e., for example improving parameter values. This study introduces a third approach, the inclusion of other variables and relations in the prediction model, so-called re-specification, to find a prediction model that may predict better than the previous one. Disadvantageous to model recalibration and respecification is the computational power that is needed. However, thanks to increased computational power, model recalibration has already become a common practice in forecasting (Hewamalage et al., 2023). Re-specification now becomes a more relevant option to prediction modeling because of the wide availability of millions of potential social media and or GT predictor variables that have the tendency to structurally change over time. Consequently, the selection of suitable variables emerges as a more pressing concern than further complicating forecasting models..

In our study, out-of-sample forecasts for the period October 2016 till December 2022 are produced 49 times in this study with each of the out-of-sample forecasting procedures (i.e., rolling window with model recalibration (RC), expanding window with model recalibration (EC), rolling window with model respecification (RS), and expanding window with model respecification (ES)), each time using a different window size ranging from 48 to 96. After averaging over the 49 forecasts for the out-of-sample October 2016 till December 2022, a single robust forecastis obtained for each out-of-sample forecasting procedure. For the best performing forecasting procedure, the multiple linear regression model that is used is extended with an autoregressive (AR) component to add a lagged version of the dependent variable (Hyndman & Athanasopoulos, 2018).

4. Results

The first part of this section presents the senstivity of forecasting accuracy to the window size. Subsequently, the forecasts produced forecast are evaluated . Finally, the AR(1) is added.
The out-of-sample period used for each window size is October 2016 till December 2022, while the in-sample periods vary between 1-10-2008 and 1-12-2012. Figure 1, 2, 3, and 4 give the results and indicate that the forecast error, and thus accuracy, is sensitive to the window size. However, the relationship between window size and forecast error is complex and can hardly be captured with an equation. Moreover, for each out-of-sample forecasting procedure, the relationship between window size and forecast error is different. Not only does the pattern differ, but the magnitude of the effect is also different. This has led to the decision to average over the forecasts obtained from the 49 different windows that were used in training the model.

Figure 1-4 summarize that the out-of-sample forecasting procedures relying on model re-specification result in more accurate forecasts than those relying on model recalibration both for the rolling window (Figure 1 vs 3) and the expanding window (Figure 2 vs 4).

Table 1 further explains only gives weak evidence that when model re-specification is used, the rolling window will produce more accurate results than the expanding window. Given that the error diagnostics of the best performing forecasting procedure (RS with window size of 48) indicated that some information is missing, an autoregressive component with a lag order of 1 is added, referred to as AR(1). Using an AR(1) component means that last month's youth unemployment rate is used to predict this month's youth unemployment rate. The forecast is now produced for a larger out-of-sample than previously used, ranging from October 2012 till December 2022.

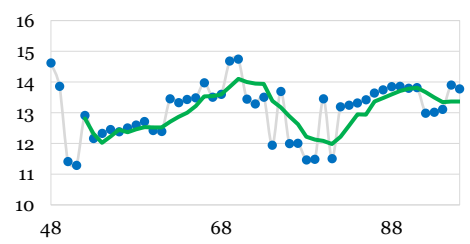


Figure 1. Forecast error rolling re-calibration to window size

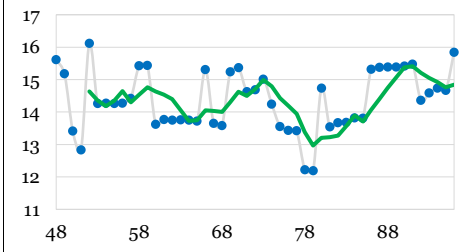


Figure 2. Expanding calibration forecast error

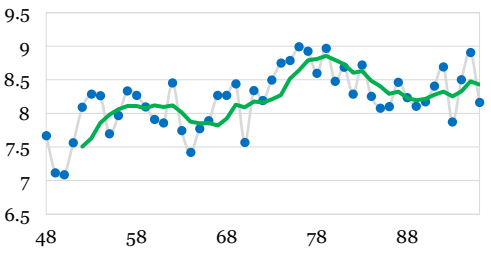


Figure 3. Forecast error rolling window re-specification

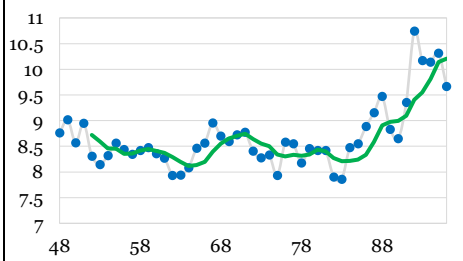


Figure 4. Forecasting error expanding window re-specification

Note: MAPE is presented on the y-axis. The MAPE for a certain window size is calculated from a sample of 75 forecasts and their corresponding errors. Window size is presented on the x-axis. The green line in the graphs represents the moving average of the MAPE. thick line concerns moving average.

The autocorrelation of the forecast errors is reduced when the AR(1) component is added to the multiple linear regression model, resulting in no significant autocorrelation at any lag. When the AR(1) component is added, the forecasting errors are also less dispersed and lower than when the AR(1) component is not added. The assumptions of multiple linear regression are fulfilled to a greater extent when an AR(1) component is used in addition to the GT variables. The one month ahead forecasts of the youth unemployment rate are more accurate when

supplementing the multiple linear regression model with an AR(1) component, with MSE being .4375 as opposed to .8079. Similarly, RMSE is substantially down from .8988 to just .6615. Finally, MAPE is also lower when the AR(1) component is used in addition to the GT variables, decreasing from 6.75% to just 4.73%.

Figure 5 shows that the forecasts align well with the actual youth unemployment rate, even for the period characterized by COVID-19. The overall fit of the forecasting model with reality has an Adjusted R-Squared of 91.33%. Although the multiple regression model, relying on the RS procedure for out-of-sample forecasting, is relatively unbiased and accurate this section revealed that it is important to supplement GT data with additional data, like the AR(1) component.

5. Conclusions and Discussion

This study reminds forecasting literature of the limitations inherent to GT data. Moreover, this study found that forecasts relying on solely GT data are substantially improved when additional information, like an autoregressive component, is added. The findings also contribute to literature by: (1) Raising awareness on the importance of picking the correct out-of-sample forecasting procedure, and (2) demonstrating that forecasts can be improved by using a different out-of-sample forecasting procedure. The results of our study also aligns with Shen et al. (2020), revealing that a larger window size may lead to lower forecasting accuracy, both for the rolling window and the expanding window.

Table 1. Error metrics

	MSE (n=75)	RMSE (n=75)	MAPE (n=75)
ES	.7302	.8545	7.807%
EC	1.936	1.391	13.34%
RS	.7816	.8841	7.213%
RC	1.589	1.260	11.94%

Note: One month ahead forecasts averaged over windows from 48 till 96 months.

We find that model re-specification substantially improves the accuracy when out-of-sample forecasting the youth unemployment rate. Compared to model recalibration, model re-specification yields 44% more accurate forecasts of the youth unemployment rate. This finding is supported with 99% of confidence. Consequently, the dominance of model re-specification, as opposed to model recalibration may be generalizable for both the rolling window and the expanding window procedure for out-of-sample forecasting.

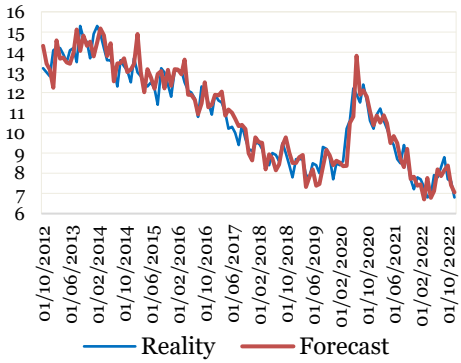


Figure 5. Best youth unemployment GT forecasts aligned with reality & AR(1) added. Note: Rolling window size of 48.

We acknowledge some limitations in our work. First, model re-specification requires large computational power, especially when more complex modelling techniques are used. For example, an autoregressive integrated moving average model takes around 60 times longer to process out-of-sample forecasts than a multiple linear regression model when using model respecification. Second, this study merely established correlation, and not causation, between GT data and the youth unemployment rate. It could be the case that the youth unemployment rate is explaining GT data more strongly than the other way around. Third, the keywords that are used could be subject to noise. For example, keywords like "werkloosheid" (unemployment) do not solely reflect searches done by the unemployed. Rather, searches for this keyword could simply reflect an interest in the current state of the economy. Fourth, this study introduced coverage bias by using merely Dutch keywords, therefore excluding individuals that don't speak Dutch. Fifth, this study only forecasted one month ahead, limiting the practical value.

Inspired by the limitations, there are various future research suggestions. Specifically we suggest that the selection of keywords/variables could be fully automated. For example, ChatGPT could be prompted to return keywords related to a certain domain ontology. Subsequently, Google Tends data for these keywords could be obtained automatically.

References

Behnen, P., Kessler, R., Kruse, F., Gómez, J. M., Schoenmakers, J., & Zerr, S. (2020). Experimental Evaluation of Scale, and Patterns of Systematic Inconsistencies in Google Trends Data. *Communications in Computer and Information Science*, *1323*, 374–384. https://doi.org/10.1007/978-3-030-65965-3_25/TABLES/4

Borup, D., & Schütte, E. C. M. (2020). In Search of a Job: Forecasting Employment Growth Using Google Trends. *Journal of Business & Economic Statistics*, *40*(1), 186–200. https://doi.org/10.1080/07350015.2020.1791133

CBS. (2023). *Arbeidsdeelname en werkloosheid per maand*. https://opendata.cbs.nl/#/CBS/nl/dataset/80590ned/table?dl=770B2

Cebrián, E., & Domenech, J. (2022). Is Google Trends a quality data source? *Applied Economics Letters*, *30*(6), 811–815. https://doi.org/10.1080/13504851.2021.2023088

Chatfield, C., & Xing, H. (2019). *The analysis of time series: an introduction with R* (7th ed.). CRC Press.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record*, *88*(special issue SI), 2–9.

D'Amuri, F., & Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting*, *33*(4), 801–816. https://doi.org/10.1016/j.ijforecast.2017.03.004

Eichenauer, V. Z., Indergand, R., Martínez, I. Z., & Sax, C. (2021). Obtaining consistent time series from Google Trends. *Economic Inquiry*, *60*(2), 694–705. https://doi.org/10.1111/ecin.13049

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012–1014. https://doi.org/10.1038/nature07634

Hewamalage, H., Ackermann, K., & Bergmeir, C. (2023). Forecast evaluation for data scientists: common pitfalls and best practices. *Data Mining and Knowledge Discovery*, *37*(2), 788–832. https://doi.org/10.1007/s10618-022-00894-5

Hyndman, R., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.com.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, *343*(6176), 1203–1205. https://doi.org/10.1126/science.1248506

Li, X., Shang, W., Wang, S., & Ma, J. (2015). A MIDAS modelling framework for Chinese inflation index forecast incorporating Google search data. *Electronic Commerce Research and Applications*, *14*(2), 112–125. https://doi.org/10.1016/j.elerap.2015.01.001

McLaren, N., & Shanbhogue, R. (2011). Using internet search data as economic indicators. *Bank of England Quarterly Bulletin*, *51*(2), 134–140. http://econpapers.repec.org/RePEc:boe:qbullt:0052

Mulero, R., & García-Hiernaux, A. (2021). Forecasting Spanish unemployment with Google Trends and dimension reduction techniques. *SERIEs*, *12*(3), 329–349. https://doi.org/10.1007/s13209-021-00231-x

Naccarato, A., Falorsi, S., Loriga, S., & Pierini, A. (2018). Combining official and Google Trends data to forecast the Italian youth unemployment rate. *Technological Forecasting and Social Change*, *130*, 114–122. https://doi.org/10.1016/j.techfore.2017.11.022

Shen, Z., Zhang, Y., Lu, J., Xu, J., & Xiao, G. (2020). A novel time series forecasting model with deep learning. *Neurocomputing*, *396*, 302–313. https://doi.org/10.1016/j.neucom.2018.12.084

Simionescu, M., & Cifuentes-Faura, J. (2022). Can unemployment forecasts based on Google Trends help government design better policies? An investigation based on Spain and Portugal. *Journal of Policy Modeling*, *44*(1), 1–21. https://doi.org/10.1016/j.jpolmod.2021.09.011

Singhania, R., & Kundu, S. (2021). Forecasting the United States Unemployment Rate by Using Recurrent Neural Networks with Google Trends Data. *International Journal of Trade, Economics and Finance*, *11*(6). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3801209

Te Brake, G. (2017). *Unemployment? Google it! Analyzing the usability of Google queries in order to predict unemployment*. Universitat de Barcelona.

Tran, U. S., Andel, R., Niederkrotenthaler, T., Till, B., Ajdacic-Gross, V., & Voracek, M. (2017). Low validity of Google Trends for behavioral forecasting of national suicide rates. *PloS One*, *12*(8), e0183149–e0183149. https://doi.org/10.1371/journal.pone.0183149

Tuhkuri, J. (2016). *Forecasting unemployment with google searches* (35). https://www.econstor.eu/handle/10419/201250

Varian, H. R. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, *28*(2), 3–27.

Vicente, M. R., López-Menéndez, A. J., & Pérez, R. (2015). Forecasting unemployment with internet search data: Does it help to improve predictions when job destruction is skyrocketing? *Technological Forecasting and Social Change*, *92*, 132–139. https://doi.org/10.1016/j.techfore.2014.12.005

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17250

Contemporary issues in Financial Technology: the role of the Internet

Daniel Broby
Department of Accounting, Finance and Economic, Ulster University, Belfast, United Kingdom.

How to cite: Broby, D. 2024. Contemporary issues in Financial Technology: the role of the Internet. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17250

Abstract

This paper investigates contemporary issues in financial technology (fintech). These are classified into six broad areas covering (1) disruption, (2) digital payments, (3) decentralization, (4) artificial intelligence, (5) open finance, and (6) financial inclusion. They are then critiqued in the context of the role of the Internet in financial services. The development and ontology of fintech is discussed, alongside the migration of financial services to the Internet. The discourse is supported by reference to the literature, as relates to the philosophical, academic, practical, and legal aspects of the issues. The paper's contribution is in providing thoughtful insight into current events and trends, and six key questions which can help to deepen our understanding of fintech and the issues surrounding it.

Keywords: *Finance; Fintech; Internet; Contemporary issues; Disruption.*

1. Introduction

This paper critiques and evaluates contemporary issues in financial technology (fintech). Specifically, how the Internet relates to, and is shaping, the global marketplace in international financial services (Lewan (2018); Economides et al. (2001)). In this respect, the definition of a contemporary issue encompasses the implications, applicability, relevance, significance, and effect of the adoption of innovative financial technology in capital markets. By focusing a scholarly lens on such issues, the paper contributes to a discourse on new ideas and perspectives in respect of the role of the Internet in finance.

The premise is that one has to understand the role that the Internet plays in financial intermediation. This is so as to conceptualize the issues that arise through fintech adoption. In this respect, the Internet can be used to match lenders with savers, and investors with investments. The Internet, as a system of interconnected computer nodes, is disintermediating

centralized markets.[1] As a result, mediators are becoming less important. This impacts both the way financial services are delivered, and the way customers experience their interaction with financial institutions (Hawkins, Mansell, and Steinmueller (1999).

Thakor (2020) defines fintech as the use of technology to provide new and improved financial services. Although useful, the shortcoming of this definition is that it does not encompass the importance of the Internet in the evolution of fintech. Broby (2021), meanwhile, uses the more focused term ”strategic fintech”. This relates to the process of thinking, acting and influencing in financial services in the Internet era. It enables the ecosystem to promote the success of both organizations and society.

Understanding the implications of this migration to the Internet is important. It goes beyond the democratization that decentralized finance (DeFi) can achieve. It is enabling financial service innovation in respect of the use of mobile devices, social networks, robotic automation, crypto assets, blockchain, distributed ledgers, and cloud computing. It is driving what Gomber, Kauffman, Parker, and Weber (2018) call the ”fintech revolution”. Through these medium, financial services are being delivered securely online, and in a more bespoke fashion.

The use of the Internet is enabling faster, more secure and more efficient financial transactions and services (Liang and Chen (2009)). The way that finance is conducted physically is changing. There are many ways that this is being manifested. The first is the increasing use of mobile devices to conduct transactions. This is happening alongside the migration of traditional banks to the Internet. Daniel and Storey (1997) suggest several catalysts for this, including a desire to add value, deliver mass customisation, and to establish reputation as a digital thought leader.

The changes, prompted by the Internet, require businesses to be redesigned. O’Reilly and Finnegan (2003) argue that, in the context of finance, the Internet is making ”bricks and mortar” an operational rather than a competitive decision. Bernstein, Song, and Zheng (2008) suggest that the ”clicks and mortar” model, is not a decision, but a strategic imperative. Many proponents of fintech view such nuances as key to understanding the impact of the Internet (see Knorr Cetina and Bruegger (2002)).

In summary, several issues arise as a result of the greater use of Internet. These have philosophical, academic, practical and legal implications. These manifest themselves at the personal, industry and societal level, some of which are now explored.

[1] The Internet is a global network of linked computing devices that use a common communications protocol, TCP/IP (Transmission Control Protocol/ Internet Protocol). TCP/IP provides a common language for interoperation between networks,switches and routers (MacKie-Mason and Varian (1994)). These in turn use a variety of local protocols (Netware, AppleTalk, DECnet and others).

2. Method – Issue identification

The working hypothesis of this paper is that the Internet is fundamentally changing the way financial services are being conducted. These changes include, but are not limited to, inter-connectivity, access to data, privacy, the nature of transactions, the future of mediation, there benefit to society, and the value of mass customization. These are need to be conceptulized relative to contemporary isssues (Casula, Rangarajan, and Shields (2021)).

Issue identification is important because the contemporary use of fintech, and its intersection with the Internet, impact the way financial firms handle third party data, privacy and digital transactions. This in turn has further implications for the nature of financial mediation, and the very way financial services are delivered. It also has implications for professional competencies, skills and attitudes (Karkkainen et al (2017)).

Based on a broad review of the scholarly literature, the issues can be broken down into several typologies (Buchi et al. (2019)). This is an effective way to identify them, and helps with an understanding of the current state of knowledge. Whilst some financial technologies are innovative, many are now in widespread use. As such, identifying them is a moving goal post.

A combination of keyword searches and advanced search techniques were used to ensure a comprehensive set of sources. Criteria, such as the reputation of the journal, the credentials of the authors, and the rigor of the research methods, were used to determine the quality of these sources. The common denominator is its linkage to the Internet (Heng et al (2007)).

Once the literature search had identified common themes, the broad contemporary issue terms were put into the Web-based Google Trends in order to see how they were trending over time. Google Trends is considered a reliable tool for predicting changes in issues under discussion. Scharkow and Vogelgesang (2011) argue that it is an accurate measure of the public's interest in a topic. The results of the search are shown in Table 1.

Table 1. This shows the Google Trends search terms weekly worldwide results 01/01/2017 - 01/01/2024. A = fintech disruption; B = digital payments: ; C = decentralization; D = artificial intelligence in finance; E = open finance; F = financial inclusion.

Variable	A	B	C	D	E	F
Stan Dev	17.3	13.7	15.7	16.0	16.1	12.1
Average	48.0	70.8	60.1	49.4	65.5	60.3
High	100	100	100	100	100	100
Low	0	47	33	28	43	36

The search terms encompass many of the more granular innovations. For example, blockchain. This is being widely embraced, and fits under A, C and D in the table. The technology has the ability to support a secure transaction environment where parties operate in a trustless

environment (Shin (2019)). That said, its integration with legacy systems, and the contemporary issues surrounding its use, need to be explored more by scholars.

3. The key issues

Identifying the key issues will never be definitive (Suyano et al (2020). However, exploring them leads to a more informed debate. Also, several aspects of fintech and the Internet can be controversial. For example, disruption to established financial norms can result in job losses, and a reshaping of the status quo. The transformation of digital banking, lending, and investment platforms raises valid consumer protection concerns.

One of the challenges of the migration to the Internet lies in the vast amounts of personal and financial data. In this respect, privacy and ownership concerns add a layer of complexity to the ethical dimensions of fintech practices. As such, there is a need to strike a balance between innovation and safeguarding user information.

Figure 1 illustrates a mind map encapsulating the identified contemporary issues. This includes the worldwide impact of fintech disruption, the pervasive influence of digital payments, the implications of decentralization, the integration of artificial intelligence in finance, the emergence of open finance models, and efforts towards financial inclusion. Addressing these challenges necessitates collaboration among fintech entities, regulators, and stakeholders to foster responsible and sustainable development.

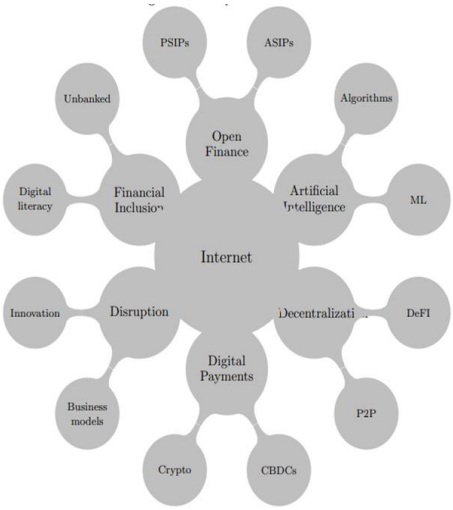


Figure 1. Depicts how the internet is at the centre of the six identified comtemporary issues in Financial Technology. The key themes branch out into sub themes.

3.1. Disruption

The Internet is a known disruptor. In the context of fintech innovation, disruption refers to its ability to impact and disintermediate existing financial institutions and business models, and the need for a proactive response from incumbent financial service providers to adapt to the changing landscape. Contemporary issues that arise from disruption relate to the direction of innovation and the resulting change to business models.

The role of disruption is articulated by the concept of disruptive innovation theory. This theory posits that the introduction of new products or services can revolutionize markets, rendering existing ones obsolete and propelling technological and economic progress. Firms failing to innovate risk obsolescence, as encapsulated in the "innovator's dilemma". Theory suggests that disruptive innovations targets underserved market segments, before expanding into mainstream markets and displacing incumbents.

Within fintech, the Internet is the catalyst for innovation, reshaping distribution methods, data access, payment forms, and data analysis. In turn, this leads to the disruption of traditional financial models. Fintech's potential is due to its capacity to transform financial institutions and business models. Its disruptive potential prompts the need for new regulations to govern the responsible and ethical use of technology, which gives rise to the following question: ***Q1. How do we navigate the legal challenges associated with fintech innovation?***

3.2. Digital payments

The Internet also provides a global platform that enables seamless and instantaneous transactions across borders. It allows for real-time payment processing. As such, security is the paramount concern. Allied to this is the reliability of the infrastructure and its resilience. There are now a whole host of innovative ways to make digital payments. These include the use of mobile wallets, blockchain payment platforms, and cryptocurrencies. Traditional payments are also undergoing transformation due to QR code payments, contactless cards, and online purchases.

Digital payments encompass digital instructions, central bank digital currencies (CBDCs), stablecoins, and cryptocurrencies (Broby (2022)). The rise of these alternatives to fiat money has implications for the financial system, financial stability and payments. The Technology Acceptance Model (TAM), the Unified Theory of Acceptance (UTA) and the Use of Technology (UTAUT) can be used to explain their adoption. These do not, however, answer the fundamental question of: **Q2.** ***Who will issue digital money in the future and how will it be structured?***

3.3. Decentralization

The role of decentralization in democratizing finance, and the potential for fintech to lead to a new financial architecture, is one of the most hotly discussed contemporary issues. Two strands of thought exist regarding DeFi. One perspective foresees a non-custodial, permission-less financial architecture, while the other emphasizes the risks of a poorly regulated ecosystem susceptible to financial crime. Scholars like Harvey et al (2021) maintain a balanced outlook, acknowledging the optimism surrounding blockchain-enabled DeFi, but highlight challenges such as the risks of centralized control, limited access, inefficiency, lack of interoperability, and opacity.

From a technological standpoint, developing blockchain and distributed ledger technologies has advantages. In the practical context, DeFi empowers users with greater control over their financial assets and data, opening avenues for a wider range of financial services. This may contribute to a more inclusive and democratized financial system. Concerns include the potential for illegal activities, jurisdictional ambiguity, smart contract liability, cybersecurity risks, and taxation issues. That said, DeFi represents the possibility to establish a more transparent, open, and accessible financial system. The question is: **Q3** ***How will DeFi co-exist alongside traditional centralized marketplaces?***

3.4. Open finance

Open finance leverages Application Programming Interfaces (APIs) to enable the secure and seamless sharing of financial data. The contemporary issue relates to the potential for this to disrupt existing financial institutions through the provision of Banking as a Service (BaaS). There is also the need for a proactive response from traditional financial firms to adapt.

Open finance is driven by the philosophy of promoting transparency and accessibility. The underlying principle is to provide greater control over their financial data, facilitating easier access and sharing. It promotes increased competition, innovation, and consumer choice as a result. The issue is whether this will create a more personalized approach to financial services by allowing companies to offer more targeted and relevant products and services.

Incumbent financial institutions need to respond to the changing landscape. This requires investing in new technologies and designing new business models. It involves collaborating with third-party data providers, and fostering a culture of innovation and experimentation. Needless to say, such actions are not without cost, and therefore debated. The question is: ***Q4. What will be the impact of open finance on banking, and what are the associated risks and challenges?***

3.5. Financial inclusion

There are a many contemporary issues related to financial inclusion. Contemporary questions arise around sustainable finance, inclusiveness and impact investing. Also, around the potential for fintech to help drive positive social and environmental change.

The "beggar on the street" problem represents the potential negative impact of financial technology and digital technology on individuals unable to access or use these technologies. If financial services are primarily available through digital channels, those without internet or mobile device access may be excluded. As a result, laws and regulations need to be well defined in order to determine access to financial services. Fintech, through technology, can enhance financial inclusion. For example, by using mobile and online platforms to reach remote or underserved areas. Also, employing artificial intelligence and machine learning can improve the accessibility of financial services.

Fintech contributes to financial literacy by providing accessible and engaging tools for understanding and managing finances. It can also support the development of socially and environmentally responsible financial products and services. Used for good, it can facilitate impact investing, provide access to microfinance, and develop platforms for sustainable finance. The question is: ***Q5. How do we ensure inclusivity in a world where not everyone has access to the Internet?***

3.6. Artificial Intelligence

The use of artificial intelligence (AI) and machine learning in finance, mass customization, and the potential impacts on jobs and the economy. A contemporary debate exists regarding consciousness, human thought, and the limitations of artificial systems. Some contend that despite AI's remarkable capabilities, it lacks essential human qualities such as subjective experience, self-awareness, and genuine understanding.

AI's potential to revolutionize finance is acknowledged by all. It enables computers to make complex financial decisions, automate processes, and handle challenging tasks. However, ethical considerations arise, questioning the replacement of human decision-makers. The employment implications, potential biases, and transparency also need to be considered, as does the risk of unintended consequences. Liability and privacy concerns arise from the vast data processed by AI systems, prompting the question: ***Q6. How is individual data protected when used to generate tailored financial solutions?***

4. Conclusion

In conclusion, there are many contemporary issues in financial technology. The Internet is reshaping the global marketplace and the delivery of financial services. It is resulting in a

process of disintermediation. Understanding these issues contributes to a deeper understanding and enhanced scholarly discourse.

Six broad contemporary issue classifications are identified based on an aggregation of Internet related fintech dynamics. These include (1) disruption, (2) digital payments, (3) decentralization, (4) artificial intelligence, (5) open finance and, (6) financial inclusion. Within each, there are several underlying contemporary issues. Six questions derived from these focus on (1) legal challenges, (2) issuance of money, (3) the role of central marketplaces, (4) impact on banking, (5) inclusivity and (6) data and privacy.

In summary, the paper contributes to the discourse on new ideas and perspectives regarding the role of the Internet in finance and provides valuable insights to inform policy decisions and improve society. It highlights the many ways in which the Internet is enabling financial service innovation and changing the way finance is conducted physically. However, the changes prompted by the Internet also require businesses to be redesigned. This has philosophical, academic, practical, and legal implications. These manifest themselves at the personal, industry, and societal level.

References

Bernstein, F., Song, J.-S., & Zheng, X. (2008). "bricks-and-mortar" vs."clicks-and-mortar": An equilibrium analysis. *European Journal of Operational Research*, 187 (3), 671–690.

Broby, D. (2021). Financial technology and the future of banking. *Financial Innovation*, 7 (1).

Broby, D. (2022). Central bank digital currencies: policy implications. *Law and Financial Markets Review*, 16(1-2), pp.100-115.

Buchi, G., Cugno, M., Luca, F., Zerbetto, A., and Castagnoli, R., (2019). New banks in the 4th industrial revolution: A review and typology. *In Proceedings of 22nd excellence in services international conference* (pp. 1–21).

Casula, M., Rangarajan, N., & Shields, P. (2021). The potential of working hypotheses for deductive exploratory research. *Quality & Quantity*, 55 (5), 1703–1725.

Daniel, E., & Storey, C. (1997). On-line banking: strategic and management challenges. *Long Range Planning*, 30 (6), 890–898.

Economides, N., et al. (2001). The impact of the internet on financial markets. *Journal of Financial Transformation*, 1 (1), 8–13.

Gomber, P., Kauffman, R. J., Parker, C., & Weber, B. W. (2018). On the fintech revolution: Interpreting the forces of innovation, disruption, and transformation in financial services.

Harvey, C. R., Ramachandran, A., & Santoro, J. (2021). Defi and the future of finance. *John Wiley & Sons.*

Hawkins, R., Mansell, R., & Steinmueller, W. E. (1999). Toward digital intermediation in the information society. *Journal of Economic Issues*, 33 (2), 383–391.

Heng, S., Meyer, T., & Stobbe, A. (2007). Implications of web 2.0 for financial institutions: Be a driver, not a passenger. *Deutsche Bank Research, E-conomics*, 63 .

Karkkainen, T., Panos, G. A., Broby, D., & Bracciali, A. (2017). On the educational curriculum in finance and technology. In *International conference on internet science* (pp. 7–20).

Katsh, E. (1993). Law in a digital world: Computer networks and cyberspace. Vill. L. Rev., 38, 403.

Knorr Cetina, K., & Bruegger, U. (2002). Global microstructures: The virtual societies of financial markets. *American Journal of Sociology*, 107 (4), 905–950.

Lewan, M. (2018). The Internet as an enabler of FinTech. In The rise and development of Fintech (pp. 190-204). *Routledge*.

Liang, C.-J., & Chen, H.-J. (2009). How to lengthen, deepen and broaden customer–firm relationships with online financial services? *Journal of Financial Services Marketing*, 14 (3), 218–231.

O'Reilly, P., & Finnegan, P. (2003). Internet banking systems: an exploration of contemporary issues. *Journal of systems and information technology*.

Scharkow, M., & Vogelgesang, J. (2011). Measuring the public agenda using search engine queries. *International Journal of Public Opinion Research*, 23 (1), 104–113.

Suryono, R.R., Budi, I. and Purwandari, B., 2020. Challenges and trends of financial technology (Fintech): a systematic literature review. *Information*, 11(12), p.590.

Thakor, A. V. (2020). Fintech and banking: What do we know? *Journal of Financial Intermediation*, 41 , 100833.

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17781

The potential of Google Trend in estimating the absorption rate of European structural funds

Nicola Caravaggio[1], **Eleonora Pierucci**[2], **Giuliano Resce**[1]

[1]Department of Economics, University of Molise, Campobasso, Italy, [2]Department of Economics, Roma Tre University, Rome, Italy.

How to cite: Caravaggio, N.; Pierucci, E.; Resce, G. 2024. The potential of Google Trend in estimating the absorption rate of European structural funds. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17781

Abstract

This study investigates the relationship between Google Trends (GT) interest in European Structural and Investment Funds (ESIF) and the absorption rate across 27 European Union countries. Utilizing a two-way fixed effect methodology, we analyse annual GT data from 2007 to 2016. Results reveal a consistently positive and statistically significant explicative power of lagged values of GT interest on absorption rates. The findings suggest that the online search behaviour regarding ESIF correlates with fund absorption, revealing the potential predictive value of GT data in the context of regional cohesion policies. This study contributes to the literature on the practical applications of GT across diverse domains and underscores its relevance in predicting the implementation of EU cohesion policies.

Keywords: *Google Trends, European Structural and Investment Funds, Cohesion Policy*

1. Introduction

Dealing with regional imbalances has consistently proven to be a challenging task, leading different entities, such as the European Union, to formulate extensive frameworks for cohesion policies over time (Farole et al., 2011). The cohesion policy is the EU's primary investment policy: a substantial portion of its activities and budget is devoted to reducing the gap between regions, with a specific focus on rural areas, regions undergoing industrial transition, and those facing natural or demographic disadvantages (European Commission, 2022). For the programming period 2014-2020, it involved over 350 billion euros (32.5% of the overall EU budget), with approximately 200 billion allocated to the European Regional Development Fund (ERDF), around 83 billion to the European Social Fund (ESF), and about 63 billion to the Cohesion Fund (European Parliament, 2024).

With substantial funds at their disposal, a key focus of cohesion policy in member states is to strengthen their ability to utilize allocated funds, achieve higher absorption rates, and enhance overall effectiveness. Despite efforts to address economic and territorial disparities, some countries, especially new members, struggle with low absorption rates (Incaltarau et al., 2020). In this regard, recent studies explore the role of administrative capacity, political governance, and other factors in explaining these differences (Surubaru, 2017; Incaltarau et al., 2020; Cunico et al., 2022). This evidence shows that improving government effectiveness and combating corruption are crucial for successful fund absorption.

The literature mentioned earlier does not delve into the unexplored aspect of potential resistance among local residents to participate in funding opportunities. In practical terms, when both citizens and policymakers show disinterest in these funding opportunities, they opt not to engage in calls, and they do not exert pressure on the local administration to participate either. This paper explores the possibility of analyzing this research question leveraging Google Trends (GT), which is a real-time daily and weekly index of the volume of queries that users enter Google. The idea is surprisingly simple: people interested in the EU cohesion found topic tend to search information about it. The literature has largely shown that Google Trend data are often correlated with various economic indicators, and it has been shown to be helpful for short-term economic prediction (Choi, Varian, 2012).

Results show that the GT interest for European Structural and Investments Funds (ESIF) has a positive impact on the absorption rate of European (27) countries, especially for lagged values, stressing how this could potentially represent a predictive tool for the allocation of European funds.

2. Implementations of GT data in the literature

Scholars increasingly turn to GT across diverse contexts (Jun et al., 2018) starting from the pioneering work of Ginsberg et al. (2009) which showcased how GT could effectively monitor and forecast the progression of influenza ahead of the official reports from the Centers for Disease Control and Prevention (CDC) in the United States. Another groundbreaking contribution by Choi and Varian (2009) elucidates how GT data aids in forecasting initial claims for unemployment benefits in the United States. More recently, Niesert et al. (2020) demonstrate the out-of-sample predictive capabilities of GT for unemployment but not for consumer confidence and consumer price index in US, UK, Canada, Germany, and Japan. Therefore, the author's interpretation is that online search could represent a reliable gauge of individuals' personal situation, but less reliable for lesser "personal" variables. Furthermore, by analyzing unemployment insurance initial claims in US, Borup et al. (2023) demonstrate the out-of-sample potential of GT in predicting weekly initial claims which emerges to be strongly linked to the Covid-19 crisis. The recent worldwide pandemic inspired several works which investigated the

association between Covid cases and specific GT queries (Kornellia & Syakurah, 2023) such as for the US case by Kurian et al. (2020) and Liu et al. (2022). Still within the US case, Lampos et al. (2021) and demonstrated the importance of GT in predicting Covid cases and deaths more than two weeks in advance.

In the realm of retail, automotive, and home sales, Choi and Varian (2012) delved into the efficacy of seasonal autoregressive models and fixed-effects models incorporating pertinent GT variables. Their findings underscored the superior performance of models integrating these predictors compared to those excluding them. Within the financial domain, various studies harnessed GT data to predict stock market trends using neural networks (Hu et al., 2018; Fan et al., 2021) or to identify "early warning signs" in financial markets (Preis et al., 2013; Petropoulos et al., 2022). Yu et al. (2019) demonstrate how GT data is useful in forecasting global oil consumption. Starting from a study of the cointegration and Granger causality test between these data, they show how their data-drive forecasting analysis with GT data improved traditional techniques carried out without relying on this new data source.

Eichenauer et al. (2021) developed an R package able to construct daily GT long-run frequency-consistent indices by aggregating different search terms. They constructed this index for three German-speaking countries stressing how their indices were highly correlated with traditional economic indicators. Therefore, their analysis could represent a tool to nowcast aggregated economic indicators generally released with several weeks of delay by national statistics offices. The same objective was pursued by Woloszko (2023) which constructed an OECD weekly tracker for 46 countries which yield to real-time weekly estimates of GDP by relying on high-frequency GT data. Moreover, Kohns and Bhattacharjee (2023) implemented a Bayesian structural time series model to nowcast US real GDP growth through GT data showing how large dimensional set of searches effectively is suitable for nowcast macroeconomic data before if become effectively available. Within the works with the goal to forecast GDP growth through GT data we can include even the one of Bantis et al. (2023) focused on Brazil and US.

Furthermore, the extensive application of GT data extends to epidemiology, serving as a real-time surveillance tool for influenza (Broniatowski et al., 2013). Additionally, GT data proves invaluable beyond predictive domains, aiding in user geolocation of Twitter (now X) data (Zola et al., 2020). Nevertheless, these approaches and the broader utilization of GT are not immune to critiques and limitations, hence their implementation in analysis should be cautious (Cook et al., 2011; Lazer et al., 2014; Nagao et al., 2019).

3. Data and methodology

In our analysis we aim to estimate the absorption rate of all European Structural and Investments Funds (ESIF) based on GT interest over time for 27 EU countries. We constructed our database by retrieving absorption rate data from the SF 2007-2013 Funds Absorption Rate of the

European Commission (2024) which represented our response variable. For the construction of our core determinant, the GT interest, we retrieved monthly country-level time series data from January 1, 2007, up to December 31, 2023, for the macro topic "European Structural and Investments Funds". We decided to start our time series from 2007 for two major reasons: (i) in this year the first cycle (2007-2013) of ESF started; (ii) Eichenauer et al. (2021) selected the same date to start their GT analysis since it is the year when the first iPhone was introduced, an event that represented a kind of shock which relentlessly boosted the use of mobile internet. In Figure 1 we reported the google trend interest at worldwide level and for Italy and Greece over the considered period.

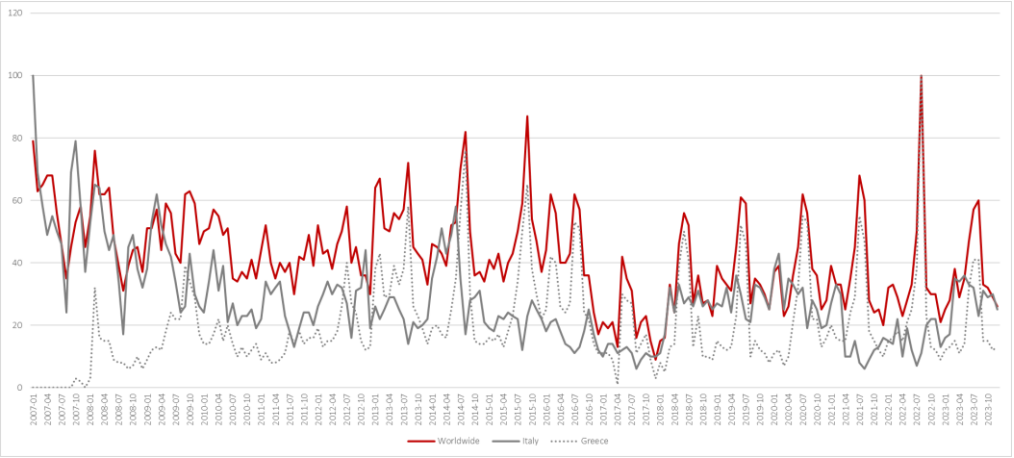


Figure 1 Google Trend interest for the topic European Structural and Investments Funds (January 1, 2007 - December 31, 2023).

GT data was aggregated (averaged) on a year basis and weighted through year-basis cross-country interest for the same topic. This weighting procedure was necessary to make data comparable across country since for each query (country-time frame) automatically GT rescale data from 0 to 100. Moreover, to account for cross-country differences in terms of internet use, we further multiply each country-year data for the corresponding inverse value of internet penetration. The final database has a panel structure with a time coverage which spans from 2007 up to 2016. The necessity to shorten the considered years was dependent on the reliability of absorption rate data.

We implemented a two-way fixed effect methodology (Baltagi, 2021) with Driscoll and Kraay (1998) robust standard errors, able to account for cross-sectional dependency, heteroscedasticity, and serial correlation (Hoechle, 2007). Results are reported on TABLE. We performed seven models characterized by different combinations of control variables. In the first model (1), the simplest one, we included only GT interest (esif_w_int) as predictor of the absorption rate. In the second model the lagged the GT interest variable while the in the third

one (3) we included both the simultaneous and lagged term. This dual analysis has been applied even in models 4 and 5 where we included some additional control variables, which have been delayed as well in models 6 and 7. The control variables included in the model are: gdp per capita (gdp_cap), the deficit and the debt share of GDP (respectively, deficit_gdp and debt_gdp), a proxy for institutional variables (iqi_avg) constructed as average of the five institutional quality index of the Worldwide Governance Indicators (WB, 2024b) , population density (pop_den), and share of rural population (pop_rur). We retrieved these additional variables from the World Development Indicators of the WB (2024a) apart from the two target variables of the European Stability and Growth Pact (GDP), whose data were retrieved from Eurostat (2024), instead.

4. Results and discussion

Results, reported in Figure 2, show a positive and statistically positive explanatory power of the GT interest for ESIF on the absorption rate. Despite hardly quantifiable, when the GT index increases by one point, the relative absorption rate increases by an amount of percentage points which ranges from 0.017 to 0.032. This first empirical exercise highlight that the GT index can elucidate absorption rates. Its significance primarily stems from past values of the GT index, indicating its ability to forecast the effective implementation of policies through the absorption of distributed funds. Given the growing accessibility of big data and the necessity for monitoring indicators that policymakers can employ to assess the ongoing effects of policies, GT data emerges as a promising predictor for overseeing the implementation of European regional policies.

	1	2	3	4	5	6	7
esif_w_int	0.0281***		0.0115	0.0222**	0.0115		0.00642
	(0.008)		(0.015)	(0.009)	(0.010)		(0.012)
L.esif_w_int		0.0320***	0.0295***		0.0248***	0.0186**	0.0173**
		(0.007)	(0.006)		(0.006)	(0.008)	(0.007)
gdp_cap				-0.000518***	-0.000569***		
				(0.000)	(0.000)		
deficit_gdp				-0.160*	-0.0971*		
				(0.074)	(0.045)		
debt_gdp				-0.00901	-0.0179		
				(0.023)	(0.025)		
iqi_avg				2.825	3.589		
				(4.105)	(4.173)		
pop_den				0.000308	0.0123		
				(0.022)	(0.022)		
pop_rur				-0.428	-0.648		
				(0.474)	(0.589)		
L.gdp_cap						-0.000617***	-0.000615***
						(0.000)	(0.000)
L.deficit_gdp						-0.234**	-0.234**
						(0.080)	(0.079)
L.debt_gdp						-0.0464*	-0.0464*
						(0.022)	(0.022)
L.iqi_avg						-3.397	-3.354
						(5.138)	(5.230)
L.pop_den						-0.000604	-0.000526
						(0.026)	(0.025)
L.pop_rur						-0.460	-0.461
						(0.503)	(0.504)
Observations	270	243	243	270	243	243	243
R-squared	0.976	0.975	0.975	0.977	0.976	0.976	0.976
Adjusted R-squared	0.972	0.971	0.971	0.972	0.971	0.972	0.971

Standard errors in parentheses
* p<0.1, ** p<0.05, *** p<0.01

Figure 2 Estimation of the absorption rate in UE 27 countries (2007-2016).

References

Baltagi, B. (2021). *Econometric Analysis of Panel Data*. Springer.

Bantis, E., Clements, M. P., & Urquhart, A. (2023). Forecasting GDP growth rates in the United States and Brazil using Google Trends. *International Journal of Forecasting, 39*(4), 1909-1924.

Borup, D., Rapach, D. E., & Schütte, E. C. M. (2023). Mixed-frequency machine learning: Nowcasting and backcasting weekly initial claims with daily internet search volume data. *International Journal of Forecasting, 39*(3), 1122-1144.

Broniatowski, D. A., Paul, M. J., & Dredze, M. (2013). National and local influenza surveillance through twitter: an analysis of the 2012-2013 influenza epidemic. *PloS one, 8*(12), e83672.

Choi, H., & Varian, H. (2009). Predicting initial claims for unemployment benefits. Google Inc, 1 (2009), 1–5.

Choi, H., & Varian, H. (2012). Predicting the present with google trends. *Economic record, 88*, 2–9.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. Economic record, 88, 2-9.

Cook, S., Conrad, C., Fowlkes, A. L., & Mohebbi, M. H. (2011). Assessing google flu trends performance in the United States during the 2009 influenza virus a (h1n1) pandemic. *PloS one, 6*(8), e23610.

Correlations between covid-19 cases and google trends data in the United States: A state-by-state analysis. In Mayo clinic proceedings (Vol. 95, pp. 2370–2381).

Cunico, G., Aivazidou, E., & Mollona, E. (2022). Decision-making traps behind low regional absorption of Cohesion Policy funds. *European Policy Analysis, 8*(4), 439-466.

Driscoll, J. C., & Kraay, A. C. (1998). Consistent covariance matrix estimation with spatially dependent panel data. *Review of economics and statistics, 80*(4), 549-560.

Eichenauer, V. Z., Indergand, R., Martínez, I. Z., & Sax, C. (2022). Obtaining consistent time series from google trends. *Economic Inquiry, 60*(2), 694–705.

European Commission (2022). 8th Cohesion Report: Cohesion in Europe towards 2050. Luxembourg: Publications Office of the European Union, 2022.

European Commission (2024). SF 2007-2013 Funds Absorption Rate. Link: https://cohesiondata.ec.europa.eu/2007-2013-Finances/SF-2007-2013-Funds-Absorption-Rate/kk86-ceun/about_data [Accessed: March 7, 2024].

European Parliament (2024). Fact Sheets on the European Union - 2024. Link: https://www.europarl.europa.eu/factsheets/en/section/195/regional-and-cohesion-policy [Accessed: March 7, 2024].

Eurostat (2024). Government finance statistics and EDP statistics. Link: https://ec.europa.eu/eurostat/web/government-finance-statistics/data [Accessed: March 7, 2024].

Fan, M.-H., Chen, M.-Y., & Liao, E.-C. (2021). A deep learning approach for financial market prediction: Utilization of google trends and keywords. *Granular Computing, 6*, 207–216.

Farole, T., Rodr´ıguez-Pose, A., & Storper, M. (2011). Cohesion policy in the european union: Growth, geography, institutions. *JCMS: Journal of Common Market Studies, 49*(5), 1089–1111.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*(7232), 1012–1014.

Hoechle, D. (2007). Robust standard errors for panel regressions with cross-sectional dependence. *The stata journal, 7*(3), 281-312.

Hu, H., Tang, L., Zhang, S., & Wang, H. (2018). Predicting the direction of stock markets using optimized neural networks with google trends. *Neurocomputing, 285*, 188–195.

Incaltarau, C., Pascariu, G. C., & Surubaru, N. C. (2020). Evaluating the determinants of EU funds absorption across old and new member states–The role of administrative capacity and political governance. *JCMS: Journal of Common Market Studies, 58*(4), 941-961.

Jun, S.-P., Yoo, H. S., & Choi, S. (2018). Ten years of research change using google trends: From the perspective of big data utilizations and applications. *Technological forecasting and social change, 130*, 69–87.

Kohns, D., & Bhattacharjee, A. (2023). Nowcasting growth using google trends data: A bayesian structural time series model. *International Journal of Forecasting, 39*(3), 1384-1412.

Kornellia, E., & Syakurah, R. A. (2023). Use of google trends database during the covid-19 pandemic: systematic review. *Multidisciplinary Reviews, 6*(2), 2023017–2023017.

Kurian, S. J., Alvi, M. A., Ting, H. H., Storlie, C., Wilson, P. M., Shah, N. D., . . . others (2020).

Lampos, V., Majumder, M. S., Yom-Tov, E., Edelstein, M., Moura, S., Hamada, Y., . . . Cox, I. J. (2021). Tracking covid-19 using online search. *NPJ digital medicine, 4*(1), 17.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of google flu: traps in big data analysis. *Science, 343*(6176), 1203–1205.

Liu, Z., Jiang, Z., Kip, G., Snigdha, K., Xu, J., Wu, X., . . . Schultz, T. (2022). An infodemiological framework for tracking the spread of sars-cov-2 using integrated public data. *Pattern Recognition Letters, 158*, 133–140.

Nagao, S., Takeda, F., & Tanaka, R. (2019). Nowcasting of the us unemployment rate using google trends. *Finance Research Letters, 30*, 103–109.

Niesert, R. F., Oorschot, J. A., Veldhuisen, C. P., Brons, K., & Lange, R. J. (2020). Can Google search data help predict macroeconomic series?. *International Journal of Forecasting, 36*(3), 1163-1172.

Petropoulos, A., Siakoulis, V., Stavroulakis, E., Lazaris, P., & Vlachogiannakis, N. (2022). Employing google trends and deep learning in forecasting financial market turbulence. *Journal of Behavioral Finance, 23*(3), 353–365.

Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using google trends. *Scientific reports, 3*(1), 1684.

Surubaru, N. C. (2017). Administrative capacity or quality of political governance? EU Cohesion Policy in the new Europe, 2007–13. *Regional Studies, 51*(6), 844-856.

WB. (2024a). World Development Indicators. World Bank. Link: https://databank.worldbank.org/source/world-development-indicators [Accessed: March 7, 2024].

WB. (2024b). Worldwide Governance Indicators. World Bank. Link: https://www.worldbank.org/en/publication/worldwide-governance-indicators#home [Accessed: March 7, 2024].

Woloszko, N. (2023). Nowcasting with panels and alternative data: The OECD weekly tracker. *International Journal of Forecasting*.

Yu, L., Zhao, Y., Tang, L., & Yang, Z. (2019). Online big data-driven oil consumption forecasting with Google trends. *International Journal of Forecasting*, *35*(1), 213-223.

Zola, P., Ragno, C., & Cortez, P. (2020). A google trends spatial clustering approach for a worldwide twitter user geolocation. *Information Processing & Management*, *57*(6), 102312.

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17804

From Crisis to Opportunity: A Google Trends Analysis of Global Interest in Distance Education Tools During and Post the COVID-19 Pandemic

Priyanga Dilini Talagala[1] , **Thiyanga S. Talagala**[2]

[1]Department of Computational Mathematics, University of Moratuwa, Sri Lanka, [2]Department of Statistics, Faculty of Applied Sciences, University of Sri Jayewardenepura, Sri Lanka.

How to cite: Talagala, P. D.; Talagala, T. S. 2024. From Crisis to Opportunity: A Google Trends Analysis of Global Interest in Distance Education Tools During and Post the COVID-19 Pandemic. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17804

Abstract

This study investigated the impact of COVID-19 on global attention towards different distance education tools. We used Google Trend search queries as a proxy to quantify the popularity and public interest in different distance education solutions under 11 sub-segments, which include collaboration platforms, online proctoring, and resources for psychosocial support. The study employs both visual and analytical approaches to analyse global web search queries during and post the COVID-19 pandemic. Through cross-correlation analysis and dynamic time-warping analysis, the study confirms the contemporaneous and lead-lag relationships between COVID-19 and distance education-related search terms. Furthermore, the study highlights the critical role of psychosocial support in promoting the well-being of students and teachers during a pandemic. The study emphasizes the importance of Google footprint analysis in determining the most popular online education resources designed for different educational goals. This feature allows educators to gain insight into prominent distant education options, boosting their online teaching.

Keywords: *Online Learning; Online Teaching; Distance Education Solutions; COVID-19 Pandemic; Google Trend Search Queries; Psychosocial Support in Education*

1. Introduction

Despite a long history of substantial changes and improvements to the delivery and communication processes, COVID-19 has ushered in a new era of distance education, leading numerous educational stakeholders to take the concept seriously (Hosen, 2022; Richmond et al., 2020; Al Karim et al., 2022). The unexpected move from schools to homeschooling on a massive scale left children, educators, and parents vulnerable, resulting in millions of education-

related internet searches during the pandemic (Andrews, Richmond, & Marciano, 2021). Surprisingly, search spikes for distance education-related inquiries (Figure 1(c)) coincide with increased COVID-19 cases (Figure 1(a)) and related internet searches (Figure 1(b)), even though distance education has a longer history than the COVID-19 pandemic.

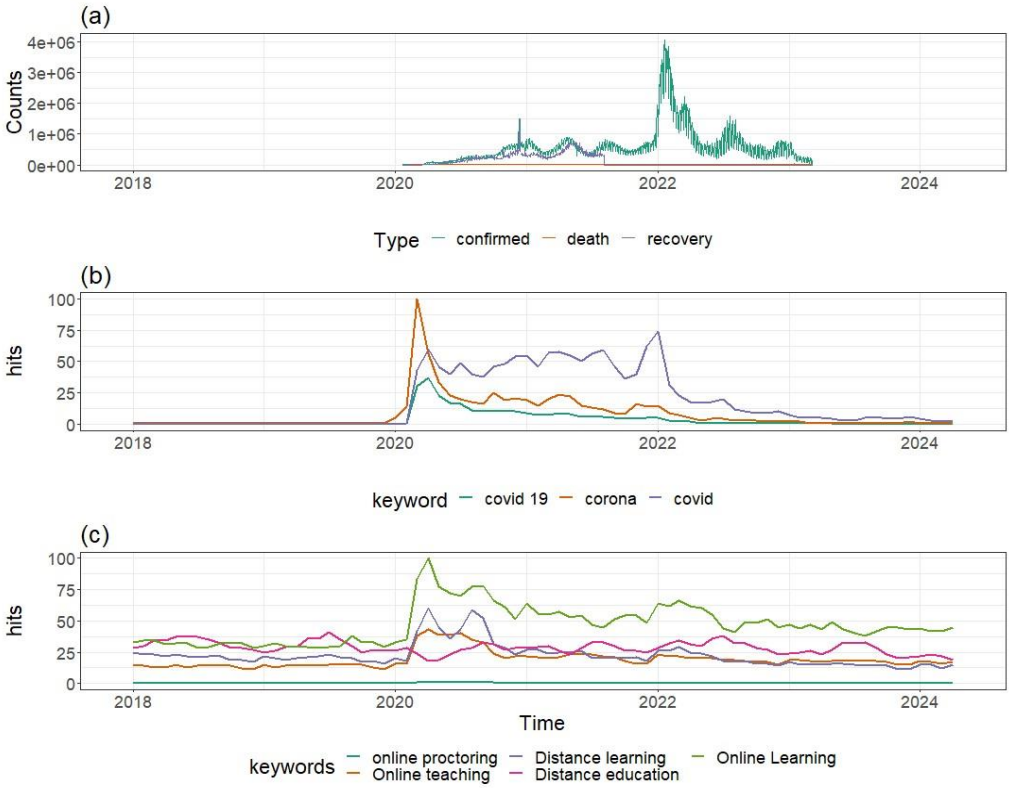


Figure 1. Visualization of data from Google search trends. (a) COVID-19 cases worldwide (b) Google search trends of COVID-19 related terms (c) Google search trends of distance education-related terms.

Google Trends, an open-source web analytics tool, offers users access to internet search data, providing valuable insights into population behaviour (Nuti et al., 2014). With ongoing discussions about the most suitable technology to support student learning, we explore whether Google Trends search queries can serve as a proxy for gauging the popularity and public interest in various distance education options. Our analysis primarily focuses on quantitative digital footprint data from December 2019 to April 2024. This study addresses two critical questions in distance education: (1) What solutions have emerged to meet the demands of distance education during and post the COVID-19 pandemic? (2) Which distance learning solutions have garnered widespread attention and public interest amid and in the aftermath of the COVID-19

pandemic? Through our examination, the resulting Google Trend footprint serves as an initial guide for identifying popular distance education tools across different educational purposes, including digital learning management systems, platforms compatible with basic mobile phones, options with robust offline functionality, Massive Open Online Course (MOOC) platforms, self-directed learning content, mobile reading applications, and collaboration platforms supporting live-video communication. This resource allows educators to streamline their search efforts and explore prominent distance education solutions to enhance their online teaching practices.

2. Methodology

Data on COVID-19 cases were collected from the coronavirus package in the R software which extracts information from the Johns Hopkins University Center for Systems Science and Engineering (JHU CCSE) data repository (Dong, Du, & Gardner, 2020). We utilized weekly Google Trends search queries as a proxy to examine the attention given to various distance learning solutions during and post the COVID-19 pandemic. All searches were conducted within the timeframe from December 1, 2019, to April 30, 2024, encompassing both the duration of the pandemic and the subsequent period.

According to Vaughan and Romero-Frías (2014), refining Google Trends searches to a specific category, such as education, enhances the relevance and accuracy of the data by minimizing noise. In this study, we concentrated on the education category. Google Trends offers relative search volume rather than absolute search numbers for a given term, adjusted based on the total searches within the specified geography and time frame (Cebrián et al., 2023). The resulting series ranges from 0 to 100, with each data point representing search interest relative to the highest point in the series for the chosen region and time. A value of 100 indicates the peak popularity of a term within the specified time frame.

Deciding on the appropriate search term in Google Trends is crucial, as each term carries different search volumes. Various strategies have been employed in previous studies to determine suitable search terms. Some studies relied on intuition or brainstorming processes for term selection (Vaughan & Romero-Frias, 2014). In our study, we considered the list of distance learning solutions published by UNESCO when choosing our search term (UNESCO, 2020a). Although these solutions are not explicitly endorsed by UNESCO, they typically possess extensive reach, a sizable user base, and evidence of effectiveness (UNESCO, 2020a). We conducted worldwide search queries for each keyword, using the keyword itself as the "search term." This approach enabled us to search for the exact text strings entered by users. However, Vaughan and Romero-Frias (2014) noted that abbreviations or acronyms generally have higher search volumes than their corresponding full names. Nonetheless, we opted not to use acronyms in our study, given our focus on specific tools and techniques available in the market to address distance education needs. Using acronyms could lead to confusion with other entities.

Our analysis encompassed 11 distinct sub-categories, including digital learning management systems, platforms designed for basic mobile phone usage, solutions with robust offline capabilities, Massive Open Online Course (MOOC) platforms, self-directed learning resources, mobile reading apps, collaboration platforms facilitating live-video communication, tools for educators to generate digital learning content, external repositories of distance learning resources, online proctoring tools, and resources aimed at offering psychosocial support. This segmentation was primarily guided by UNESCO's compilation of distance learning solutions issued in response to the COVID-19 pandemic.

In Google Trends, users are limited to searching for up to five queries simultaneously (Vaughan & Romero-Frias, 2014). To address this limitation, we conducted our analysis by entering up to five distance learning tools at a time under each segmentation and recording their relative ranking scores. Utilizing an iterative pairwise comparison method, we initially identified the series with the highest search volume during the study period. This iterative process allowed us to determine the tool with the highest relative ranking score, which served as a reference point for obtaining the relative ranking scores of other tools within the same segmentation (Vaughan & Romero-Frias, 2014).Our analysis employed both visual and analytical techniques to examine shifts in web search queries globally and identify emerging evidence regarding the impact of the COVID-19 pandemic on distance education. Furthermore, we assessed the correlation between weekly web searches and the global COVID-19 case count. Cross-correlation analysis and the dynamic time warping algorithm (Giorgino, 2009) were employed to explore the relationship and similarities between the search volume data related to COVID-19 and distance education.

3. Results and Discussion

According to Jarynowski et al. (2020), public concerns regarding various issues can follow a pattern akin to an epidemic, progressing from an initial phase of increasing interest, termed "early adoption," to a subsequent phase of widespread interest, referred to as "majority," and eventually declining in popularity, described as the "laggers stage." This life-cycle explanation of Google Trend search patterns is further supported by the observations depicted in Figures 1(b) and 1(c).

A pairwise comparison was conducted using the Dynamic Time Warping (DTW) algorithm, which involves stretching or compressing two time series locally to align them as closely as possible. The resulting distance between the two series is computed by summing the distances of individual aligned elements (Giorgino, 2009). Specifically, two pairwise comparisons were made: "corona" with "online learning" and "covid" with "online learning," aiming to assess the impact of COVID-19 on distance education. Both comparisons revealed significant areas of overlap and exhibited similar dynamic patterns. The DTW analysis yielded normalized

distances of 0.1326197 and 0.1828299 for the "corona" with "online learning" and "covid" with "online learning" comparisons, respectively. A normalized distance closer to 0 indicates a higher degree of similarity between the sequences, while a value closer to 1 suggests greater dissimilarity.

Furthermore, the close relationship between COVID-19 related search patterns and the search patterns for terms such as "Online learning," "Distance learning," and "Online teaching" was clearly established through time series cross-correlation analysis. Notably, the significant cross-correlation coefficients at lag zero provide strong evidence confirming the contemporaneous relationship between the search terms associated with COVID-19 and those pertaining to distance learning.The UNESCO-published list offers diverse distance education solutions for various educational needs. Figures 2 and 3 present Google trend footprint, reflecting global attention towards these solutions and aiding in identifying popular tools. However, the Google Trends series over multiple panels is incomparable because they indicate "relative" search volume for given terms. They are comparable only within a given panel.

Figure 2. Google trend footprint analysis of distance learning solutions during and post the COVID-19 pandemic.

Figure 3. Google trend footprint analysis of distance learning solutions during and post the COVID-19 pandemic.

During the COVID-19 pandemic, Google Classroom gained widespread attention for its user-friendly features. However, public interest in systems designed for basic mobile phones is

notably limited, punctuated only by occasional, significant spikes. Kolibri garnered attention for its offline functionality, crucial for learners in underserved areas. Canvas stood out among MOOC platforms, while Quizlet excelled in self-directed learning. Reads emerged as a preferred mobile reading app, and WhatsApp surpassed Zoom in live-video communication platforms. Collaboration tools like Nearpod and EdPuzzle also gained traction. External repositories like Brookings and Education.Nation received significant attention for their extensive resources. Online proctoring tools such as Pearson VUE saw increased interest amid the pandemic. Peaks in March 2020 aligned with school closures (UNESCO, 2020b), reflecting the urgency for remote learning solutions, while spikes in August 2020 indicated ongoing challenges in reopening schools fully (UNESCO, 2020b).

Effective psychosocial assistance is critical in improving the mental health and overall well-being of both students and teachers during a pandemic. This is evident from the considerable jumps in Google search volume index graphs for the phrase 'Psychosocial support' during academic breaks in the third quarter of 2020, 2021, 2022 and 2023 (Figure 4), particularly in the 'Education' category. The unexpected spike in the first quarter of 2022, diverging from the typical third-quarter patterns, aligns with the global peak of COVID-19 cases recorded during that same period. This surge underscores the profound impact of the pandemic on education, resulting in widespread partial school closures due to the significant disruptions to conventional onsite classroom teaching. These surges highlight the compelling need to address and resolve these challenges. Analyzing secondary data in this manner is critical, as these characteristics may be overlooked due to decreased engagement and communication between students and educators during school closures.

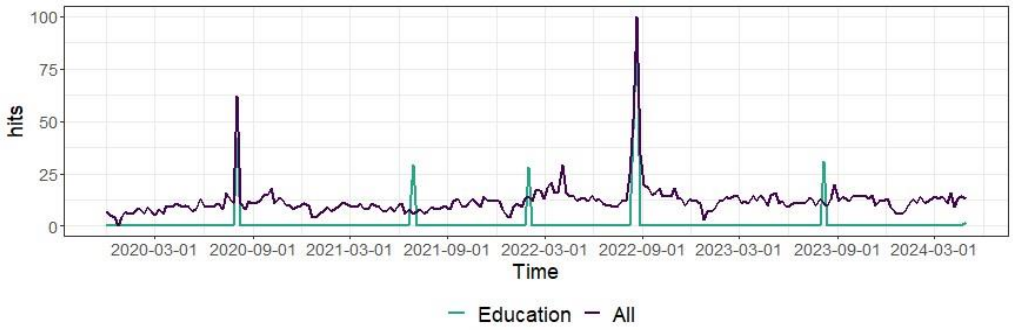


Figure 4: Google search trend of the term psychosocial support under 'All' categories and 'Education' category during and post the COVID-19 pandemic.

4. Conclusions and Further Work

The primary aim of this research was to explore the utility of Google Trend search queries in gauging the popularity and public interest surrounding various distance education solutions. This study represents a pioneering effort, being the first comprehensive analysis of web search

behaviour concerning distance learning solutions during and post the COVID-19 pandemic. While prior studies have examined Google Trend search queries related to education, they have typically focused on a limited number of search terms or different research objectives. In contrast, our investigation extends this research by analyzing a broader array of relevant search terms and conducting a more detailed examination of the popularity and interest in distance learning solutions amid the pandemic.

Our findings underscore the substantial impact of the COVID-19 pandemic on global interest in distance education, as evidenced by the strong correlation between COVID-19-related search terms and those pertaining to distance education in Google Trend queries. Moreover, our thorough analysis of Google Trend data reveals a surge in the popularity and public attention toward diverse distance learning solutions during this period. These insights offer valuable guidance for educators seeking to navigate the myriad options available in the market and identify the most effective tools for different educational needs. Furthermore, our study highlights the crucial role of learning tools in enhancing student engagement in online learning, a point emphasized in previous research. Consequently, our findings hold particular relevance for developers and educational institutions, providing them with valuable information to identify competitors in the market and enhance existing tools. Additionally, the challenges posed by the high costs of distance learning solutions, inadequate financial support, and limited understanding of available options during and post the COVID-19 pandemic underscore the importance of our study's insights. By offering valuable guidance to stakeholders in education—including teachers, administrators, policymakers, and students—our research facilitates informed decision-making with limited time and effort.

Moreover, our study emphasizes the critical importance of psychosocial support in promoting the mental health and well-being of both students and teachers during a pandemic, as evidenced by spikes in search volume during academic breaks and school closures. Addressing these overlooked aspects through effective measures is paramount, as demonstrated by our analysis of secondary data, which enables the identification and prioritization of areas requiring attention. By prioritizing psychosocial support, educational institutions can contribute to fostering a healthier and more resilient school community.

While our analysis of Google Trends data serves as a valuable proxy for quantifying the popularity and public interest in distance education solutions during the pandemic, it is essential to acknowledge that popularity alone does not guarantee quality. Future research endeavors should focus on evaluating the efficiency and effectiveness of the most popular tools to ensure engaging and effective online learning experiences. Additionally, considering the limitations of Google Trends data—including its capture of search behaviour from a subset of the population with internet access and its relative volume representation—further investigation is warranted at the national and regional levels to identify specific measures for ensuring the quality of distance education on a broader scale.

Furthermore, as highlighted by Cebrián et al. (2023), there is a need for future research to delve deeper into the data quality aspects of Google Trends search queries. While our study has provided valuable insights into the popularity and public interest in distance education solutions during and post the COVID-19 pandemic, there remains a gap in understanding the accuracy, reliability and coverage of Google Trends data. Exploring the scope and determinants of potential inaccuracies in Google Trends data across various contexts would be instrumental in enhancing its utility for educational research and decision-making.

Acknowledgment

This work was supported by UNESCO and the International Development Research Centre, Ottawa, Canada.The views expressed herein do not necessarily represent those of UNESCO, IDRC or its Board of Governor.

References

Al Karim, M., Masnad, M. M., Ara, M., Rasel, M., & Nandi, D. (2022). A Comprehensive Study to Investigate Student Performance in Online Education during Covid-19. *International Journal of Modern Education & Computer Science*, *14*(3).

Carter Andrews, D. J., Richmond, G., & Marciano, J. E. (2021). The teacher support imperative: Teacher education and the pedagogy of connection. *Journal of teacher education*, *72*(3), 267-270.

Cebrián, E., & Domenech, J. (2023). Is Google Trends a quality data source?. *Applied Economics Letters*, *30*(6), 811-815.

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases*, *20*(5), 533-534.

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: the dtw package. *Journal of Statistical Software* **31**, 1–24. 10.18637/jss.v031.i07

Hosen, M. B. (2022). Impact of Telecommunication Service Quality in Bangladesh on Online Education during Covid-19. *International Journal of Modern Education and Computer Science*, 14(6), 65–75.

Jarynowski, A., Wójta-Kempa, M., & Belik, V. (2020). Perception of emergent epidemic of COVID-2019/SARS CoV-2 on the Polish Internet. *arXiv preprint arXiv:2004.00005*.

Nuti, S. V., Wayda, B., Ranasinghe, I., Wang, S., Dreyer, R. P., Chen, S. I., & Murugiah, K. (2014). The use of google trends in health care research: a systematic review. *PloS one*, *9*(10), e109583.

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Richmond, G., Cho, C., Gallagher, H. A., He, Y., & Petchauer, E. (2020). The critical need for pause in the COVID-19 era. *Journal of Teacher Education*, *71*(4), 375-378.

UNESCO. (2020a). Distance learning solutions. Retrieved from https://en.unesco.org/covid19/educationresponse/solutions

UNESCO. (2020b). Adverse consequences of school closures. Retrieved from https://en.unesco.org/covid19/educationresponse/consequenc

Vaughan, L., & Romero-Frías, E. (2014). Web search volume as a predictor of academic fame: An exploration of G oogle trends. *Journal of the Association for Information Science and Technology*, *65*(4), 707-720.

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17794

A Methodological Framework for Examining Sociotechnical Imaginaries during the implementation of emerging technologies

Suania Acampa
Southern Centre for Digital Transformation; Department of Social Sciences, University of Naples Federico II, Italy.

How to cite: Acampa, S. 2024. A Methodological Framework for Examining Sociotechnical Imaginaries during the implementation of emerging technologies. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17794

Abstract

In the social sciences, there is a growing interest in how people imagine and interpret the impact of technology, captured by the term "Sociotechnical Imaginaries." This concept highlight society's collective tech expectations can influence policies, investments, and mobilizing resources. The study proposes a methodological approach to analyse these imaginaries based on a digital, mixed and sequential design with different phases: it starts by reviewing European documents on new technologies to identify regulatory models among countries. Then, it involves discussions with experts across sectors to understand their views on technology and its societal implications. The final goal is to create an Opinion Dictionary via hybrid Opinion Mining, providing a tool for innovation researchers to map out narratives central to technological futures. This methodology aims to uncover the narratives driving society's technological expectations and their transformation processes, offering valuable insights for policymakers and analysts.

Keywords: *Sociotechnical imaginaries, Narratives Analysis, Emerging Technologies, Digital Methods.*

1. Introduction

In the field of social sciences, there has been an increasing focus on exploring how social actors perceive and anticipate technological innovations. Jasanoff and Kim (2015) introduced the notion of "Sociotechnical Imaginaries," which are collectively held visions and expectations regarding technology's future role and impact within a society. Beginning with a contemplation of current technological states, these scenarios lay out a horizon of expectations that, while not guaranteed to be fulfilled, are nonetheless capable of mobilizing resources currently, legitimizing political decisions, guiding strategic investments in specific sectors, and initiating

social dynamics. Thus, the capacity of social actors to envisage the technological future is not merely theoretical but also has performative aspects (Borup et al., 2006). The examination of narratives has recently zeroed in on their role in fostering projectivity (Andersen et al., 2020) and their utility in probing imaginaries concerning the technological future. Narratives about the future can vary depending on the temporal aspect (e.g., they might lead to alternating cycles of excitement and disillusionment), the type of technology and the change it promises, and the communities of actors engaged in the discursive practice of sharing or contesting expectations about the future (Borup et al., 2006). Social actors share divergent narratives, form distinct expectations about the technological future, and endeavor to translate these expectations into conceivable imaginaries (Mager and Katzencach, 2021). Therefore, in an era marked by rapid technological and social changes, grasping the future's imaginaries through narratives is crucial for anticipating potential evolutionary scenarios and steering public policies. This underlines the necessity for an analytical approach that leverages discursive practices to delve into expectations around emerging technologies and digital transformation processes. The research presented herein seeks to bridge this gap to develop an analytical framework capable of semi-automatically reconstructing the narratives that underpin imaginaries on the technological and digital future, to comprehend how future expectations can direct change processes.

2. Methodological Path and Analysis Phases

This study falls within the Science and Technology Studies (STS) domain, specifically within the STS branch that focuses on the construction processes of sociotechnical scenarios. By "emerging technologies," we refer to technological innovations at an early development stage with significant potential to profoundly impact society, the Economy, and culture. These technologies typically exhibit rapid changes, high uncertainty, and often significant ambiguity regarding their future applications and implications. The concept of emerging technology is dynamic, evolving as innovations surface and others mature to become integrated into daily life (Rotolo et al., 2015). The research addresses the following questions in detail: Q1 - What kind of sociotechnical future is envisioned in narratives about technologies? Q1- Which socio-technical imaginaries are developed about emerging technologies? Q2 - Which governance and transformation processes do these relate to? Q3 - Which actors are involved? Q4 - How do the imaginaries among different actors and in relation to different technologies vary? To address these questions and meet the overall goal of developing an analytical approach that leverages discursive practices for exploring expectations concerning emerging technologies and digital transformation processes, the research was structured around a sequential exploratory methodological approach employing digital and mixed methods. This approach is divided into four phases summarized in Figure 1.

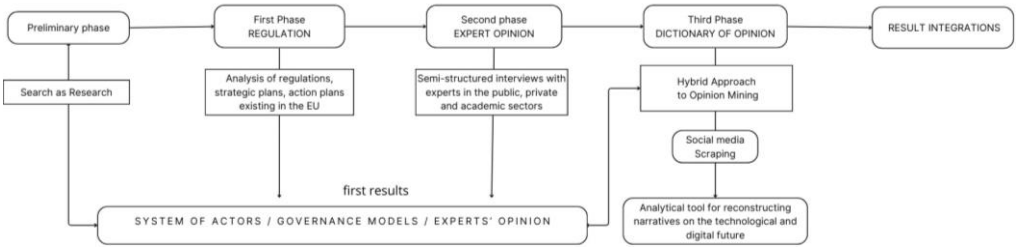


Figure 1: Research Design

At this point we will describe all the phases of the research and illustrate the proposed analytical model. Being a work in progress, we are currently working on the second analysis phase, the results of which will be presented at the conference.

2.1. Preliminary Phase. Defining the Investigation Field

This phase is grounded in the "Search as Research" approach (Salganik, 2019), which posits that online search activities, including search queries and navigation patterns, can serve as valuable data sources for digital social research. This methodology was employed to establish the parameters and boundaries of the research, trace trajectories of interest, and define the field of investigation. This phase aimed to delineate the research context, identify the technologies to be analyzed, pinpoint European countries that have adopted or initiated regulations, strategies, or action plans concerning these technologies, set the time frame, identify data collection sources, and determine search keywords. The selected time frame encompasses the last five years, during which the EU has increasingly engaged in consultations on emerging technologies. The technologies identified as fitting the "emerging technologies" category include Artificial Intelligence (AI), Robotics, High-Performance Computing (HPC), Digital Twins, and Automated Decision Support (ADS). As an initial step for each technology, it was necessary to identify the online sources for gathering existing documentation under investigation. Beginning at the European level, all online portals provided by the European Community offices and the European Parliament were identified, offering official and comprehensive access to legal and legislative documents, reports on action strategies, and investments by the EU and its member countries. For some identified technologies, EU portals are exclusively dedicated to them (e.g., for AI), featuring national chapters for each member state (plus Norway and Switzerland, but excluding the UK) with detailed and updated factsheets on adopted policies and respective national strategies. Each national factsheet includes information from the Organisation for Economic Co-operation and Development (OECD), which systematically collects and presents data on the regulatory strategies of countries worldwide, including policies, governance guidelines, and financial support. The OECD.AI dashboards also offered valuable guidance for navigating the national portals of the United Kingdom, which are excluded from the European

chapters The OECD sources were also a valid support for orienting oneself among the national portals of the United Kingdom, excluded from the European factsheets. Ultimately, the European Commission and the OECD provide general guidelines for governments that still need to publish such documents. For each identified technology, keywords were generated (in appendix) to initiate web search queries on national ministerial sites. To build the list of keywords, Google Trends was a valid support; it has been used as a source of information for new topics of interest concerning emerging technologies. OECD dashboards have also been useful: these use real-time data to show timely trends on where, how and at what pace new technologies are being developed and used, and in which sectors. National strategies for implementing quantum technologies were encountered during the search for regulations/action plans on HPC, leading to their inclusion among the technologies under investigation.

2.2. Frist Phase. Existing Regulations in Europe Regarding Emerging Technologies

The first phase of analysis involves collecting and examining existing official documents on regulations, national strategies, and action plans across various European countries. This aims to identify guidelines, actors involved, and governance practices at the individual country and European community levels. The goal is to define standards, procedures, and guidelines for implementing and using technologies. This process highlights common and divergent positions useful for identifying policy models that guide change and strategic planning. Generally, ethical and social considerations are integrated into policies, strategies, and regulations, emphasizing the importance of responsibly managing emerging technologies. Governance elements—including ethics, privacy, security, monitoring, data management, research and development, innovation, inclusion, and professional training—may vary in scope and nature depending on national priorities, available resources, and the socio-political context. By comparing the policies and development plans of different member states, we expect to discern specific guidelines and practices adopted at both the national and European levels, offering a clear picture of how emerging technologies are managed. This comparison aims to highlight the similarities and differences in each country's approach to technological advancement, enhancing understanding of the current context to assess future trends and directions in technology regulation that drive change processes. Identifying specific strategic and regulatory models characterizing groups of countries positions the subsequent focus on Italy as a pilot investigation that could be expanded to other European nations, considering the emerged models. Based on the keywords and government sites identified in the previous analysis phase, we proceeded with the document collection. On many of the websites consulted, it was possible to conduct a thorough search using various filters, including the period, general theme (education, Economy, innovation, healthcare, etc.), and document type (this filter generally helped to exclude press reviews and journalistic articles from search results). In total, 165 documents were collected as follows: 66 documents relating to AI, 19 documents relating to Quantum Technologies, 18 documents relating to HPC, 14 documents relating to Robotics, 12 documents relating to Digital

Twins, and 36 documents relating to legislation national of the Member States that have implemented the GDPR, 9 of which have provisions regarding Automated Decision-Making Systems (referred to as art. 15 and Article 22 of the GDPR). For some countries, some documents were found to be updated versions of the same implementation plan; in this case the most recent document was considered (table 1).

What emerges is that we cannot yet talk about "regulation" about these technologies but more generally about "national action plans", "strategic plans", "development plans", etc. In the context of the development and implementation of emerging technologies at the national level, these terms describe documents and initiatives that guide a country's approach towards adopting the technology and its regulation. Although the two terms seem similar, it is possible to highlight slightly different purposes and contents: a National Strategy establishes a country's long-term vision regarding the development and use of technology. A national action plan (or agenda) is a more detailed document that sets out specific steps (how and when), initiatives, projects and resource allocations to achieve the objectives outlined in the national strategy. As is partly evident from the table, almost all EU countries have a national implementation plan for AI and Quantum Technologies. For HPC, most European countries refer to the EuroHPC program, and the search for national documentation certainly shows the presence of infrastructures in the area (super computers) but not national strategies or actual regulations; rather these documents concern financing plans or project tenders. For robotics, regulatory research, strategies and action plans almost always lead back to AI regulations. In fact, as stated in the document European Civil Law Rules In Robotics of the European Parliament Committee on Legal Affairs (2016), artificial intelligence is considered as an underlying component of "intelligent autonomous robots" (ibid. pp. 11-21) and therefore thought of as something that allows the autonomy of robotic technological systems and not something different from them. Similarly, for Digital Twins, national documents refer to data protection laws starting from the GDPR (General Data Protection Regulation), and subsequent implementations. “Digital twins” as digital representations of physical systems, processes or services are governed by the same laws that govern the use of digital data, especially those related to privacy protection and security. The European Union legislation intended to protect the personal data and privacy of individuals within the EU and the European Economic Area. Also for automated decision-making systems (ADS/ADM) the applicable general principles are to be found in the national provisions of the Member States that have accepted the GDPR on data protection and also regulated automated decision-making processes (therefore prohibitions, exceptions and guarantees). The GDPR is in fact an important reference when it comes to automated decision-making and profiling processes as well as adequate guarantees to protect people from this type of processing. The art. 22, paragraph 2, letter. b explicitly refers to any form of processing of personal data that uses automated systems, including algorithms and machine learning models, without human

Table 1. Technologies' documents collected for each European country (plus Norway and the United Kingdom)

	AI	Quantuum Tecnologies	HPC	Robotics	Digital Twins	National Provisions on art 22 GDPR (ADS)
EU	**X**	**X**	**X**	**X**	**X**	**X**
Austria	X	X				X
Belgium	X					X
Bulgaria	X					
Cyprus	X					
Croatia						
Denmark	X	X		X		
Estonia	X					
Finland	X	X				
France	X	X	X	X		X
Germany	X	X				X
Greece						
Ireland	X	X				X
Italy	X	X				
Latvia	X					
Lithuania	X					
Luxembourg	X					
Malta	X					
Norway	X		X			
Holland	X	X				X
Poland	X					
Portugal	X		X			
United Kingdom	X	X		X	X	X
Czech Republic	X					
Romania	X					
Slovakia	X					
Slovenia	X					X
Spain	X		X	X		
Sweden	X	X				
Hungary	X					X

intervention. The GDPR Europeanising data protection but delegating significant power to Member States to shape the regulatory landscape within their jurisdiction (Mayer-Schönberger and Padova 2016, p. 325). Given this initial evidence, to meet the research objectives, the documents will be analyzed through Content Analysis forms built starting from European documentation.

2.3. Second Phase. Focus on Italy: The Experts' Opinions

Through semi-structured interviews, the goal of this research phase is to engage in in-depth discussions with experts on the various technologies identified, aiming to understand their technological expectations, change processes, and the broader impact these technologies may have on the future of Italian society. Experts will be selected from the private sector, public sector, and Italian academia, considering their different roles, including managerial and technical positions. The process will begin by identifying leading companies (both private and public) in the sector and university departments dedicated to researching the technologies of interest. From this, the initial group of interviewees will be formed, consisting of academics and managers. The process will then proceed with snowball sampling, asking them to facilitate contact with technicians (in companies), and other professionals or experts with different perspectives, backgrounds, or specializations (in academia). This approach aims to understand how experts envision the future of various technologies in the country, offering an in-depth view of the future technological and digital landscape in Italy. This provides crucial insights to comprehend and navigate a rapidly evolving context. It will enable the identification of emerging trends and technological development processes that experts believe could significantly impact the future. The interviews are expected to shed light on the challenges and opportunities Italy might face in the technological realm and discussions on the ethical, social, and cultural implications of new technologies. An important aspect could be comparing expectations of what technology can achieve and practical realities, such as technological limitations and available resources, potentially leading to policy recommendations. Ultimately, the interviews should also offer valuable insights into how experts communicate the complexity of the investigated technologies to non-experts.

2.4. Third Phase. Construction of an opinion dictionary

Once the governance policies, the system of actors involved and the opinions of experts have been identified, data will be extracted from social media and processed through Natural Language Processing (NLP). Most studies using Opinion Mining techniques have focused on domains other than the one we intend to explore here; no dictionaries contain the necessary contents and concepts to process information about expectations, visions, and orientations regarding the digital and technological future. To generate empirical evidence, an opinion dictionary specific to this domain is needed. The data sample for dictionary construction will be

extracted from major social media platforms (Facebook, Twitter, TikTok) and instant messaging apps (Telegram), taking a cross-platform perspective into account to consider how the different structures of platform discussion spaces affect user experiences around shared narratives. Data will be collected using web scraping techniques based on specific keyword queries for each relevant technology, within a specific timeframe. An LDA topic modelling will be performed on the textual elements of the sample, automatically providing a topic variable for each post, and features will be extracted. The most significant words will be manually labelled based on their orientation towards the future (positive, negative, or neutral polarization), the expectations they promote, the visions they convey, and the discussed change processes. The opinion dictionary will thus include a set of words for each category. The machine learning approach will enable the labelling of the entire sample and verification of the classification's accuracy through a series of synthetic indicators. An opinion dictionary will be created through this hybrid approach (combining supervised and unsupervised methods), becoming an essential tool for reconstructing shared narratives on digital and technological transformation in the Italian context. The analytical path described is summarized in the model indicated in figure 2.

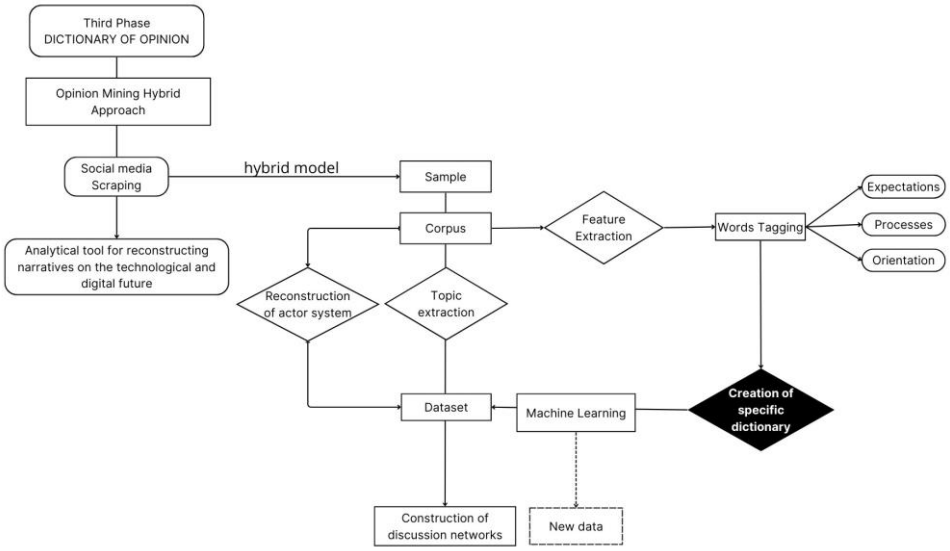


Figure 2. Model of Hybrid Approach to Opinion Mining

3. Conclusion

Through this methodological path, for each technology identified, it will be possible to: identify particular regulatory models that characterize groups of countries and this makes the subsequent focus on Italy a pilot investigation that can be extended in the future to other European nations,

taking into consideration the models that have emerged; understand how experts imagine the future of different technologies in the country, offering an in-depth view of the future technological and digital landscape in Italy and providing essential information to understand and address a rapidly evolving context; build a specific dictionary in Italian that allows the semi-automatic reconstruction of narratives on the technological future through which to investigate socio-technical imaginaries. The dictionary become a replicable tool for the semi-automatic reconstruction of narratives available for other scholars and innovation analysts.

References

Andersen, D., Ravn, S., & Thomson, R. (2020). Narrative sense-making and prospective social action: methodological challenges and new directions. *International Journal of Social Research Methodology*, 23(4), 367-375.

Borup, M.,Brown, N., Konrad, K., & Van Lente, H. (2006). The sociology of expectations in science and technology. *Technology analysis & strategic management* , 18,3 4: 285 298.

Jasanoff S, Kim SH (2015). *Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power*. Chicago: University of Chicago Press.

Mager, A., Katzenbach, C. (2021). Future imaginaries in the making and governing of digital technology: Multiple, contested, commodified. *New Media & Society* ,23,2 223 236.

Mayer-Schonberger, V., & Padova, Y. (2015). Regime change? Enabling big data through Europe's new data protection regulation. *Colum. Sci. & Tech. L. Rev.*, 17, 315.

Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? *Research policy*, 44(10), 1827-1843.

Salganik, M. J. (2019). *Bit by bit: Social research in the digital age*. Princeton University Press.

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17814

Management Accounting and Digital Technologies: A Science mapping review

Adriana Barreto[1,2], Patrícia Gomes[2], Patrícia Quesado[2], Shane O'Sullivan[1]
[1]Sustainable Development Research Institute, Technological University of the Shannon: Midlands Midwest (TUS), Ireland. [2]Research Center on Accounting and Taxation (CICF), Polytechnic Institute of Cavado and Ave (IPCA), Portugal.

How to cite: Barreto, A.; Gomes, P.; Quesado, P.; O'Sullivan, S. 2024. Management Accounting and Digital Technologies: A Science mapping review. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17814

Abstract

This study applied science mapping techniques to provide an overview of the scientific production of management accounting and digital technologies. It was considered a sample of 128 articles extracted from Scopus and WoS. Results showed that almost 80% of the articles analyzed were published in the last five years. This growth in scientific production is mainly due to technological advancements, demand for real-time information, regulatory changes and accounting standards, the need for efficiency and cost control, and the growing interest in sustainability and social responsibility issues. The conceptual structure of this sample was grouped into four clusters: digital strategy adoption, digital financial innovation, digital transformation strategies, and digital and financial sustainability. Based on the gaps identified, future research should explore management accounting and AI-based technologies within the government sector, SMEs, environmental sustainability issues, and curricular changes in university accounting courses.

Keywords: *Management accounting; digital technology; big data; artificial intellingence; bibliometric analysis; science mapping.*

1. Introduction

Management accounting has experienced a significant transformation due to the digital revolution. It has evolved from reporting historical data, to also include performance measurement and providing decision-making information (Appelbaum et al., 2017). This change is a result of several factors, including changing needs of the regulatory environment (Rautiainen et al., 2024). In this context, scientific production on Management Accounting (M.A.) and digital technologies has increased considerably in recent years, especially regarding

its transformation due to technological advancements to understand how it may improve strategic decision-making. Studies related to adopting M.A. innovation is motivated mainly due to the demand for real-time information, regulatory and accounting standards change, the need for efficiency and cost control, and growing interest in sustainability and social responsibility. Such discussions have been developed multidisciplinary, including debates related to big data, business intelligence and analytics (BI&A) tools, management control systems, Artificial Intelligence (A.I.), Machine Learning (M.L.), and blockchain.

Adopting M.A. innovation brings many benefits, such as efficiency, productivity, security and time and cost reductions (Poyda-Nosyk et al., 2023). However, the challenges are also the debated in academic literature, such as the lack of digital competencies (Steens et al., 2024) which is considered barrier to adopting digital technology for management accountants. From an organizational perspective, the lack of technological infrastructure is also considered a barrier to the adoption of these technologies (Dogru et al., 2023).

Despite the growing increase in academic production, there are still many gaps involving these issues. Thus, the purpose of this study is to understand this field through a science mapping analysis applied to a sample of (128) articles extracted from the Scopus and the Web of Science (WoS) to provide an overview of the scientific production and to map its conceptual structure. This study is relevant to understanding the multidimensional approaches regarding M.A. and digital technologies, contributing to shed light on how advanced tools are redefining its practices. This study may also contribute to academics, managers, and policymakers interested in understanding how research has been developed in this knowledge area.

2. Methods

This study adopted systematic literature review protocols, bibliometric analysis, and science mapping explore M.A. digital technologies studies indexed on the Scopus and the WoS. The combination of such multiple approaches helps the researcher to understand the topic in a more comprehensive way. Additionally, conducing a literature reviews systematically can provide quality, replicability, reliability, and validity of these reviews (Page et al., 2021; Xiao & Watson, 2019). In recent years, an increasing attention has been dedicated to the systematic study of the scientific literature, due to the availability of online databases and development of tools able to perform automatic analyses (Aria et al., 2020). Indeed, many innovations in conducting systematic reviews have emerged, including new methods (Pagani et al., 2023) and technological advances that have enabled the use of natural language processing and machine learning to identify relevant evidence (Page et al., 2021). However, researchers in business, management and related disciplines still develop cursory and narrative reviews that lack a systematic investigation of the literature developments (Linnenluecke et al., 2020).

Although systematic literature review and bibliometric analysis are different types of review methods with different purposes, as highlighted by Donthu et al. (2021), these review methods are complementary, and if integrated they may offer unique advantages in advance theory and practice in a scientific domain (Mukherjee et al., 2022). Considering this, scholars should cultivate novelty within research on business and management (Kraus et al., 2022). It includes using bibliometric analysis and science mapping techniques because it may offer unique opportunities for making a theoretical contribution, understanding their foundations, and fostering new paradigms (Paul et al., 2021; Post et al., 2020).

In general, bibliometric analysis is defined in the literature as a method applied for analyzing large volumes of scientific data to explore the evolutionary nuances and shed light on the emerging areas in a specific field (Donthu et al., 2021), based on statistical techniques (Aria et al., 2020). However, it is important to highlight that bibliometric research involves two main categories of analytical techniques: performance analysis and science mapping (Aria et al., 2020; Donthu et al., 2021; Mukherjee et al., 2022). The performance analysis is an evaluative technique for assessing productivity and impact, and science mapping is a relational technique for uncovering knowledge clusters in a field (Mukherjee et al., 2022). Thus, considering the relevance of these integrated methods to ensure a comprehensive, transparent, and replicable exploration of the literature, this study applied bibliometric performance analysis and science mapping techniques to explore the state of the art of this topic.

2.1. Research strategies

The research strategies applied included a longitudinal analysis of a bibliographic sample, the *Methodi Ordinatio* (Pagani et al., 2023), the PRISMA statement (Page et al., 2021), bibliometric analysis, and science mapping (Aria & Cuccurullo, 2017; Donthu et al., 2021; Mukherjee et al., 2022; van Eck & Waltman, 2010) to answer the following research questions:

Research Question 1: What is the current development of M.A. research on digital technologies? RQ1 aims to provide an overview of the scientific production based on bibliometric performance analysis: annual scientific production, journals, and authors.

Research Question 2: What are the main characteristics of the conceptual structure of the bibliographic sample on M.A. and digital technologies? RQ2 aims to understand this field knowledge applying science mapping techniques through term co-occurrence analysis.

2.2. Sample selection process and workflow

To answer the research questions, this study applied three main stages, as depicted in Figure 1.

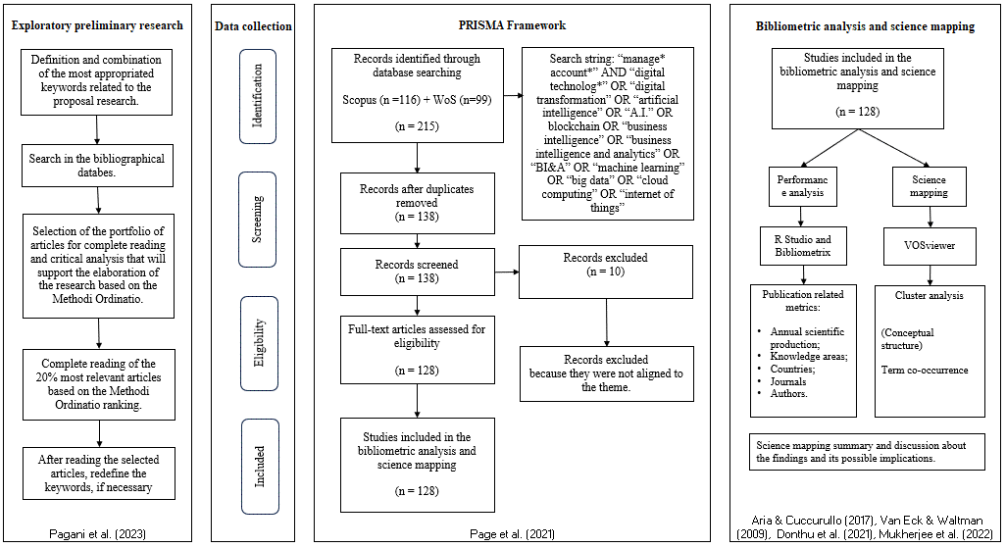


Figure 1. Scicence mapping workflow. Source: Based on Van Eck & Waltman, (2009), Aria & Cuccurullo (2017), Donthu et al. (2021), Page et al. (2021), Pagani et al. (2023).

The first phase starts with the exploratory preliminary stage to establish the subject foundation, thereby selecting the most appropriate keywords associated with the research proposal (Pagani et al., 2015; Rikhardsson & Yigitbasioglu, 2018). At this stage, the *Methodi Ordinatio*[1] was applied to systematically select articles for full reading, considering impact factor, year of publication, and number of citations (Pagani et al., 2015). The second phase included PRISMA statement to ensure a systematic search process. The query string and filters led to the retrieval of (116) articles from Scopus and (99) articles[2] from WoS, as shown in Figure 1. The same query string was applied to both databases, and the filters "Article" for document type, and "English" for language. As strategy, the authors decided not to filter by area of knowledge, considering the multidisciplinary nature of the thematic discussions identified in the initial exploratory analysis.

The next step was to export the metadata files in the BibTex file format in both databases and merge them in the RStudio using the Bibliometrix package. In this process, (77) duplicated

[1] *Methodi Ordinatio* is a methodological strategy for selecting, collecting, and ranking a bibliographic portfolio and systematically reading it (Pagani et al., 2015). This methodology aims to select and rank the papers according to their scientific relevance using three criteria in the *InOrdinatio* equation: number of citations, year of publication, and impact factor, or journal metrics (Pagani et al. 2023).

[2] Metadata retrieve reference in Scopus and WoS: February 2024.

registers were removed, resulting in a sample of (138) articles for analysis. Then, after reading all the titles and abstracts, (10) articles were excluded, considering they were not aligned with the theme, resulting in a final sample with (128) articles[3].

Then, bibliometric performance analysis and science mapping were carried out using the Bibliometrix (Aria & Cuccurullo, 2017) and the VOSviewer (van Eck & Waltman, 2010). To execute the bibliometric performance analysis, the RStudio was applied to enable visualization and analysis using the Bibliometrix. The Bibliometrix is an open-source tool R-package for performing bibliometric analyses developed by Aria & Cuccurullo (2017). The Bibliometrix package was applied to visualize information related to the evolution of scientific production, journals, and authors. The science mapping through the cluster analysis of the conceptual structure was performed using the VOSviewer, a free computer program developed by van Eck & Waltman (2010) for bibliometric maps visualization.

3. Results

3.1. Bibliometric performance analysis (RQ1)

3.1.1. Annual scientific production

Although advances in digital technologies have been considered the most important forces driving change in M.A. field, it is only recently that their impacts have attracted the attention of academics (Steens et al., 2024). This was observed in this bibliometric analysis, which showed that almost 80% of the articles analyzed were published in the last five years. Figure 2 illustrates the annual scientific production of the (128) articles, which shows this growth in recent years.

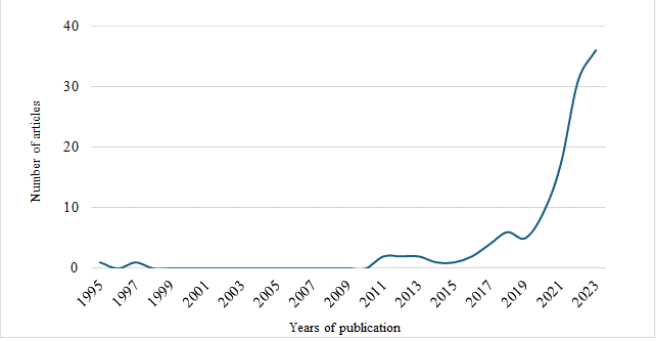


Figure 2. Annual scientifc production. Source: According to bibliographic data from Scopus and WoS.

[3] The list of 128 articles that compose the bibliographic database analyzed in this study can be accessed through the following link: https://ciencipca.ipca.pt/handle/11110/2943

The highest annual growth rates occurred in 2021 and 2022, an increase of 89% and 82%, respectively, compared to the previous years. This growth can be explained due to technological advancements (Steens et al., 2024), demand for real-time information, regulatory changes and accounting standard, need for efficiency and cost control, and growing interest in sustainability and social responsibility (Appelbaum et al., 2017; Nielsen, 2022).

3.1.2. Journals

This analysis revealed that the (128) articles were published in (91) different scientific journals. This variety of journals suggests the multidisciplinary perspectives and discussions related to M.A. and digital technologies within this sample. Despite this multidisciplinary degree, most articles were published in accounting, management, and business journals, followed by others specialising in information systems and computational-related fields. Table 1 presents the proportion of such journals, including their impact factor according to the Scimago and the Clarivate metrics. Additionally, to broaden the spectrum of analysis in relation to this other journal evaluation, the respective Academic Journal Guide (AJG) ranking, a specific journal assessment for business and management-related fields, was also provided.

Table 1. Most relevant journals. Source: Scopus, WoS, Scimago, Clarivate, and the Chartered Association of Business Schools.

Sources	2022 SJR	SJR Quartile	2022 JIF	JIF Quartile	AJG 2021	Articles	(%)
Journal of Accounting and Organizational Change	0.4	Q2	1.9	*	2	7	5.5%
International Journal of Accounting Information Systems	1.1	Q1	4.6	Q3	2	6	4.7%
Computational Intelligence and Neuroscience	*	*	*	*	*	5	3.9%
Journal of Management Control	0.7	Q2	3.3	*	2	4	3.1%
European Accounting Review	1.0	Q1	3.3	Q2	3	3	2.3%
Journal of Information Systems	1.0	Q1	1.9	Q3	1	3	2.3%
Sustainability	0.6	Q1	3.9	Q2	*	3	2.3%
Accounting, Auditing and Accountability Journal	1.7	Q1	4.2	Q2	*	2	1.6%
Accounting Education	0.7	Q1	3.2	*	2	2	1.6%
Accounting Horizons	1.0	Q1	2.5	Q3	3	2	1.6%

The analysis revealed that the *Journal of Accounting and Organizational Change* published the most articles in this sample (7). Followed by the *International Journal of Accounting Information Systems* (6). These Journals are considered relevant in the accounting field regarding issues associated with integrating accounting and information technology.

3.1.3. Authors

Figure 3 presents the top 20 productive authors on M.A. and digital technologies within this sample. The most productive author in this sample was Anca Vărzaru, with four articles published between 2022-2023, focusing on A.I. under different perspectives, such as cost, sustainability, and ethical issues. The most influential paper was *Assessing Artificial*

Intelligence Technology Acceptance in Managerial Accounting (Vărzaru, 2022), which focused in the barriers in adopting A.I. technology in M.A., such as resistance to change, organizational culture, lack of trust, and the cost of technology.

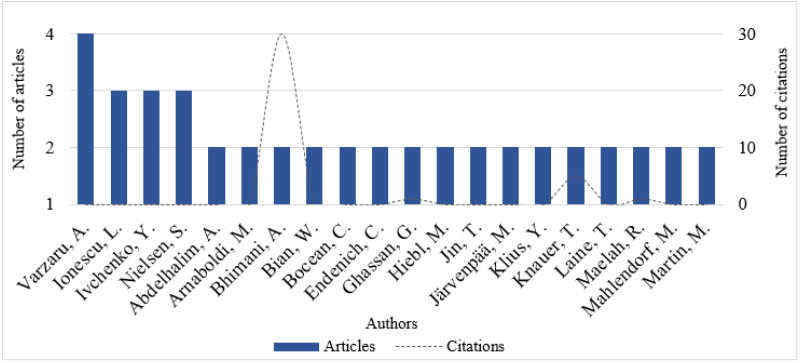


Figure 3. Most productive authors. Source: According to bibliographic data from Scopus and WoS.

Then, Luminiţa Ionescu, Yevhen Ivchenko, and Steen Nielsen that developed research related to big data processing techniques and algorithmic decision-making tools. Under the perspective of the number of citations, Alnoor Bhimani was the most cited author in *Digitisation, Big Data and the transformation of accounting information* (Bhimani & Willcocks, 2014).

3.2. Science mapping (RQ2)

3.2.1. The conceptual structure

The conceptual structure was mapped through the text-mining functionality of VOSviewer, which supports creating term maps to visualize the conceptual structure of a field (van Eck & Waltman, 2010), based on the co-occurrence of the most relevant terms within the titles and abstracts of the (128) articles. For this proposal, a full counting method and considering the minimum number of five occurrences of a term were applied in the software. Considering the algorithm default choice parameters, the 152 most relevant terms were initially selected, resulting in a sample of 71 most relevant terms after four cleaning stages to exclude terms not directly related to the topic. These terms were grouped into four clusters. Figure 4 illustrates these four clusters, revealing the four main research streams within these studies.

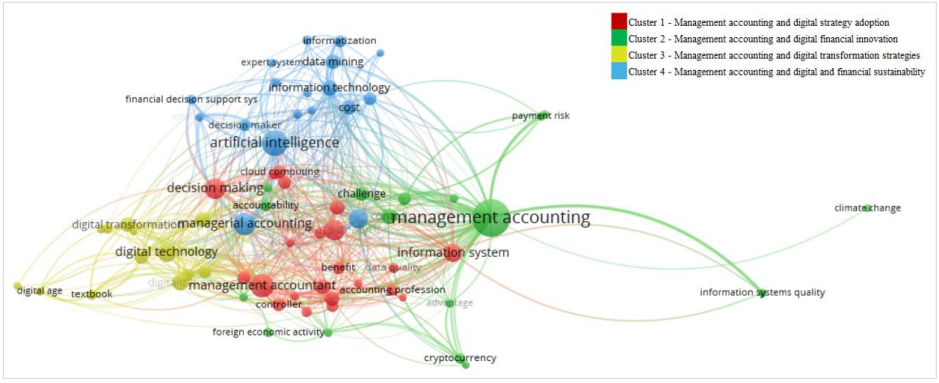


Figure 4. Term map on management accounting and digital technologies. Source: Elaborated using VOSviewer software.

Considering this analysis perspective, Table 2 presents the most relevant terms related to each cluster and their main research lines: digital strategy adoption, digital financial innovation, digital transformation strategies, and digital and financial sustainability.

Table 2. Most relevant terms in each cluster. Source: Based on Scopus and WoS, using VOSviewer.

Cluster	Cluster name	Most relevant terms
Cluster 1 - Red (22 terms)	M.A. and digital strategy adoption	Management accountant; decision making; big data analytic; information system; management accounting practice; capability; business analytic; cloud computing; controller; benefit.
Cluster 2 - Green (18 terms)	M.A. and digital financial innovation	Management accounting; control; challenge; machine learning; accountability; cryptocurrency; circular economy; foreign economic activity; payment risk; stakeholder.
Cluster 3 - Yellow (18 terms)	M.A. and digital transformation strategies	Artificial intelligence; blockchain; information technology; cost; data mining; management accounting system; informatization; decision maker; big data technology; competitive advantage.
Cluster 4 - Blue (13 terms)	M.A. and digital and financial sustainability	Digital technology; digitalization; financial accounting; sustainability; digital transformation; requirement; forecasting; digital age; economic crisis; healthcare organization.

In cluster 1 (M.A. and digital strategy adoption), the central discussion is related to the impact and benefits of adopting digital technologies, such as big data analytics, BI systems, and cloud computing, including new competencies requirements, and responsibilities. In summary, the focus of discussions in articles associated with this cluster revolves skills and capabilities necessary to extract the maximum benefit from digital technologies to subsidize the decision-making process. Some subtopics discussed in this cluster includes data quality and security.

In cluster 2 (M.A. and digital financial innovation), most discussions are related mainly to the potential and challenges of technological innovation tools in M.A., with emphasis on machine learning. Beyond expanding the efficiency and effectiveness in data evaluation processes, studies highlight the advantage of using M.L. in predictive analytics, improving capabilities for reporting and decision-making. Additionally, discussions embrace the relevance of control

mechanisms to provide accountability and effectiveness in management control, including payment risk reduction, especially within the blockchain and cryptocurrency discussions.

Cluster 3 (M.A. and digital transformation strategies) discusses mostly the importance of digital transformation strategies, emphasising changing paradigms and adapting processes due to emerging technologies, such as A.I. and blockchain. The central theme discusses how A.I. solutions can be strategically applied to create value, reduce costs, and improve competitive advantage through its potential to expand the efficiency and accuracy of accounting practices and decision-making. Blockchain technology is frequently discussed within this cluster, with studies highlighting its potential to make registration processes more transparent, safe, and efficient, ensuring control and accountability in transactions. An example of a secondary theme refers to applying this technology in the government sector and ethical issues related to A.I.

In cluster 4 (M.A. and digital and financial sustainability), most terms are associated with theoretical aspects of M.A. digitalization. Such discussions also highlight the requirements for its implementation, including the need for adaptations in financial accounting, moving it from the historical perspective practices to a more predictive perspective based on forecasting analysis to provide financial sustainability, especially during economic crises. The term "sustainability" was also applied in some studies on improving accounting information systems through new digital technologies to promote environmental sustainability practices.

5. Conclusion

This study applied science mapping techniques to explore the scientific production of M.A. and digital technologies based on a sample of (128) articles indexed in Scopus and WoS. The analysis revealed how multidisciplinary the approaches of this sample, ranging from the practical application of digital tools to issues regarding security, data quality, and the need for constant innovation and the development of skills necessary for their use. The most discussed technologies include business intelligence and analytics tools, A.I., and blockchain under different perspectives. According to the literature, the strategic adoption of digital technologies in M.A. practices produces several benefits, such as efficiency and accuracy of the information produced, supporting decision-making with more precise and real-time insights, transparency, security, and enabling better tracking and analysis of operational and financial performance.

Regarding RQ1, the analysis revealed that almost 80% of the articles analyzed were published in the last five years, due to technological advancements, demand for real-time information, regulatory changes and accounting standards, need for efficiency and cost control, and the growing interest in sustainability and social responsibility issues. The Journals that publish the most articles were the *Journal of Accounting and Organizational Change* (7) and the *International Journal of Accounting Information Systems* (6). The most productive author was Anca Vărzaru, focusing on investigations on A.I. (cost, sustainability, and ethical issues).

In response to RQ2, the science mapping revealed that the conceptual structure of this set of articles may be grouped into four main research streams: digital strategy adoption, digital financial innovation, digital transformation strategies, and digital and financial sustainability. Based on the gaps identified in this analysis, some themes are suggested as future research. The first topic suggestion concerns how digital technologies based on A.I. and blockchain can be applied in the government sector to improve financial reporting, increase the transparency of governmental management, and encourage social participation, for example, through participatory budgeting and social control mechanisms. Second, regarding SMEs, studies should research how A.I and blockchain could be applied to expand access to credit and assist managers easily and intuitively in managing financial resources. Third, it is necessary to understand how all these technological advances and A.I, combined with advances in management accounting methods, can increasingly contribute to improve social and environmental sustainability. Fourth, and no less important, studies should explore how accounting courses in universities should be adapted, emphasising statistical skills for big data analysis techniques, to encourage the minimum skills to management accountants use the potential of digital technologies, including AI and blockchain, more accurately.

This study has some limitations regarding research strategies and methods applied. Considering this, future research using other query string strategies, other bibliographic databases, and different science mapping techniques may offer additional insights.

Acknowledgements

This study was conducted at the Research Center on Accounting and Taxation (CICF) and was funded by the Portuguese Foundation for Science and Technology (FCT) through national funds (PRT/BD/154685/2023, UIDB/04043/2020, and UIDP/04043/2020). The authors thank RUN-EU and the Technological University of the Shannon: Midlands Midwest (TUS). The authors also thank the anonymous reviewers for their valuable contributions.

References

Appelbaum, D., Kogan, A., Vasarhelyi, M., & Yan, Z. (2017). Impact of business analytics and enterprise systems on managerial accounting. *International Journal of Accounting Information Systems*, *25*, 29–44. https://doi.org/10.1016/j.accinf.2017.03.003

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

Aria, M., Misuraca, M., & Spano, M. (2020). Mapping the Evolution of Social Research and Data Science on 30 Years of Social Indicators Research. *Social Indicators Research*, *149*(3), 803–831. https://doi.org/10.1007/s11205-020-02281-3

Bhimani, A., & Willcocks, L. (2014). Digitisation, Big Data and the transformation of accounting information. *Accounting and Business Research*, *44*(4), 469–490. https://doi.org/10.1080/00014788.2014.910051

Dogru, T., Line, N., Mody, M., Hanks, L., Abbott, J., Acikgoz, F., Assaf, A., Bakir, S., Berbekova, A., Bilgihan, A., Dalton, A., Erkmen, E., Geronasso, M., Gomez, D., Graves, S., Iskender, A., Ivanov, S., Kizildag, M., Lee, M., … Zhang, T. (2023). Generative Artificial Intelligence in the Hospitality and Tourism Industry: Developing a Framework for Future Research. *Journal of Hospitality and Tourism Research*. https://doi.org/10.1177/10963480231188663

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, *133*, 285–296. https://doi.org/10.1016/j.jbusres.2021.04.070

Kraus, S., Breier, M., Lim, W. M., Dabić, M., Kumar, S., Kanbach, D., Mukherjee, D., Corvello, V., Piñeiro-Chousa, J., Liguori, E., Fernandes, C., & Ferreira, J. J. (2022). Literature reviews as independent studies: guidelines for academic practice. *Review of Managerial Science*, *16*(8), 2577–2595. https://doi.org/10.1007/s11846-022-00588-8

Linnenluecke, M. K., Marrone, M., & Singh, A. K. (2020). Conducting systematic literature reviews and bibliometric analyses. *Australian Journal of Management*, *45*(2), 175–194. https://doi.org/10.1177/0312896219877678

Mukherjee, D., Lim, W. M., Kumar, S., & Donthu, N. (2022). Guidelines for advancing theory and practice through bibliometric research. *Journal of Business Research*, *148*, 101–115. https://doi.org/10.1016/j.jbusres.2022.04.042

Nielsen, S. (2022). Management accounting and the concepts of exploratory data analysis and unsupervised machine learning: a literature study and future directions. *Journal of Accounting and Organizational Change*, *18*(5), 811–853. https://doi.org/10.1108/JAOC-08-2020-0107

Pagani, R. N., Kovaleski, J. L., & Resende, L. M. (2015). Methodi Ordinatio: a proposed methodology to select and rank relevant scientific papers encompassing the impact factor, number of citation, and year of publication. *Scientometrics*, *105*(3), 2109–2135. https://doi.org/10.1007/s11192-015-1744-x

Pagani, R. N., Pedroso, B., dos Santos, C. B., Picinin, C. T., & Kovaleski, J. L. (2023). Methodi Ordinatio 2.0: revisited under statistical estimation, and presenting FInder and RankIn. *Quality and Quantity*, *57*(5), 4563–4602. https://doi.org/10.1007/s11135-022-01562-y

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Whiting, P., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *The BMJ*, *372*. https://doi.org/10.1136/bmj.n71

Paul, J., Merchant, A., Dwivedi, Y. K., & Rose, G. (2021). Writing an impactful review article: What do we know and what do we need to know? *Journal of Business Research*, *133*, 337–340. https://doi.org/10.1016/j.jbusres.2021.05.005

Post, C., Sarala, R., Gatrell, C., & Prescott, J. E. (2020). Advancing Theory with Review Articles. *Journal of Management Studies*, *57*(2), 351–376. https://doi.org/10.1111/joms.12549

Poyda-Nosyk, N., Borkovska, V., Bacho, R., Loskorikh, G., Hanusych, V., & Cherkes, R. (2023). The role of digitalization of transfer pricing in the company's management accounting system. *International Journal of Applied Economics, Finance and Accounting*, *17*(1), 176–185. https://doi.org/10.33094/ijaefa.v17i1.1096

Rautiainen, A., Scapens, R. W., Järvenpää, M., Auvinen, T., & Sajasalo, P. (2024). Towards fluid role identity of management accountants: A case study of a Finnish bank. *British Accounting Review*. https://doi.org/10.1016/j.bar.2024.101341

Rikhardsson, P., & Yigitbasioglu, O. (2018). Business intelligence & analytics in management accounting research: Status and future focus. *International Journal of Accounting Information Systems*, *29*, 37–58. https://doi.org/10.1016/j.accinf.2018.03.001

Steens, B., Bots, J., & Derks, K. (2024). Developing digital competencies of controllers: Evidence from the Netherlands. *International Journal of Accounting Information Systems*, *52*. https://doi.org/10.1016/j.accinf.2023.100667

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84(2), 523–538. https://doi.org/10.1007/s11192-009-0146-3

Vărzaru, A. A. (2022). Assessing Artificial Intelligence Technology Acceptance in Managerial Accounting. Electronics (Switzerland), 11(14). https://doi.org/10.3390/electronics11142256

Xiao, Y., & Watson, M. (2019). Guidance on Conducting a Systematic Literature Review. Journal of Planning Education and Research, 39(1), 93–112. https://doi.org/10.1177/0739456X17723971

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.18193

The Effect of Negative Emotions of Service Recipients on Negative Word of Mouth Marketing in the Health Sector

Bahar Çelik[1], Çapla Özçelik[2]

[1]Health Management Department, Kutahya Health Sciences University, Turkey, [2]Health Management Department, Kutahya Health Sciences University, Turkey.

How to cite: Çelik, B.; Özçelik, Ç. 2024. The Effect of Negative Emotions of Service Recipients on Negative Word of Mouth Marketing in the Health Sector. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.18193

Abstract

Emotional satisfaction can develop after the positive or negative emotions experienced about a purchased garment, shoe or cosmetic product are transferred to another consumer. This sharing, which can be considered as emotional satisfaction for the consumer, turns into a positive or negative purchasing experience and a positive or negative product recommendation in terms of marketing. Moreover, this situation is not only about a goods, but also about many sectors such as entertainment, health and education. For example, the positive or negative experience of a consumer who receives health care from a hospital turns into a recommendation about this hospital to their immediate surroundings. This study aims to evaluate to what extent the negative emotions of service recipients in the healthcare sector may affect negative word of mouth marketing. In this study, which is an empirical research, the survey scale developed by Wen-Hai et al. (2018) was used as the data collection method.

At the end of the study, it was observed that individuals' anger levels significantly affected their desire for revenge and negative WOM levels.

Keywords: *Marketing, Negative Word of Mouth, Negative Emotions, Health Sector.*

1. Introduction

Health services are a service that deeply affects individuals' lives and is also expensive (Martin, 2017). For this reason, competitive advantage for private hospitals and awareness studies regarding protective and preventive health services for public hospitals have gained importance. Private hospitals feel the need to create a marketing strategy on many issues such as the health services they provide, the competencies of their physicians, and the physical and technical facilities they have. Because both convincing the patient/customer group with high added value due to the cost element and managing customer relations with the understanding of lifelong

customer value can be achieved with the promotional activities of the marketing field. High healthcare costs direct patients, especially those in the low and middle income group, to public hospitals. However, providing health services in public hospitals with limited physical and human resources is becoming increasingly difficult with the increasing population. For this reason, public hospitals need social marketing activities to reduce the circulation of individuals coming to the hospital with the help of protective and preventive health services. At this stage, word of mouth marketing is a factor that strongly affects health behavior (Martin, 2017). This study aims to evaluate to what extent the negative emotions of service recipients in the healthcare sector may affect negative word of mouth marketing. Although there are some studies on word-of-mouth marketing in the health sector, gaps in this field continue (Pauli et al., 2023). Therefore, it is thought that the study will contribute to filling this gap.

2. Theoretical Review

The fact that the health sector has gone beyond compulsory treatment services and started to take on a different structure such as protective, preventive and even beauty or cosmetic services has caused hospitals to turn into service centers and patients into service customer profiles. For this reason, although it is lagging behind other sectors, the field of marketing has been integrated into the health sector and the concept of health marketing has been created. In health marketing, which is defined as a concept that enables the creation, communication and delivery of health information and interventions using customer-centered and science-based strategies to protect and improve the health of various populations, patients/consumers are segmented by market segmentation and marketing strategies specific to the target groups identified within each segment are developed (Woodside et al., 1998; CDC (Cited in Swenson et al., 2018), 2011). With the impact of digital technology, the preferences and health expectations of consumers, who have easy access to information and are becoming more aware every day, are changing. Therefore, it is important to develop marketing strategies in many health fields such as pharmaceutical companies to improve consumer attitudes and behaviors (Greenspun and Coughlin, 2012; Swenson et al., 2018).

The health sector is an important sector that is directly related to the concerns of individuals and directly affects their well-being, happiness and quality of life. Although health seems to be a simple situation that consumers can directly benefit from and will not have any problems in making decisions, Many factors such as the hospital environment, the physician's knowledge and experience, the quality of the consumables used, and the accompanying conditions make it difficult for the patient to make a decision and make the healthcare service complex. Moreover, the fact that an element such as cost is one of the leading issues in the field of health makes this situation even more complicated. Consumers can make decisions for products such as food, clothing, cosmetics and hotel management by taking into account the ratings or comments made in the digital environment. Maybe, despite all these comments and scores, the consumer may

not be satisfied with the product he bought and may pay the price for this by experiencing a bad product. However, in the health sector, which directly affects human life, the health service that the consumer receives by taking even a small risk may cost his life. For this reason, comments made by other consumers, scores or advertising studies made by the health institution are not sufficient alone in making a decision (Reinhardt, 2005; Kay, 2007). Therefore, a stronger reference such as word of mouth marketing is needed in the health sector.

Especially in the health sector, which affects human life and has a high cost and complex structure, information obtained from reference sources such as family and friends is more reliable and more effective than information obtained from advertising tools such as television, radio, newspapers and brochures, and therefore the information obtained from these sources is useful in health care. It may lead to a decrease in risk perception (Khalid et al., 2013). Because the positive or negative emotions acquired by consumers who have had similar experiences before develop the ability to make familiar choices in other consumers and cause them to exhibit similar behavior with consumers who have had this experience (Whyte, 1954; Litvin et al., 2008; Trusov et al., 2009; Chaniotakis and Lymperopoulos 2009).

Emotions that emerge as positive or negative emotions during the consumer experience can be expressed in different types such as anger, discontent, envy, worry, tense, sadness, happiness/pleased, surprise, optimism, excitement (Richins (Cited in Curwen and Park, 2014), 1997). Each emotion may appear in different ways depending on the factors affecting the service recipient before, during and after consumption. For example, after purchasing the product, the consumer may realize that there is a better alternative and regret it, the product he purchased may not meet his expectations and he may be disappointed, and the attitude of the customer representative may make the consumer angry. The consumer may feel guilty and regretful about the choices he made, or he may blame the company for the inadequacy of the service he received and feel angry. In cases where consumers have a lot to lose, such as health, the situations and consequences that occur after shopping trigger consumer emotions. When these emotions are negative, this can directly affect negative word of mouth communication. For this reason, companies need to prevent negative word of mouth communication by working towards the negative emotions experienced by the consumer after the purchasing behavior (Yi and Baumgartner, 2004; Helena Vinagre and Neves, 2008; Curwen and Park, 2014).

3. Methodology

This study aims to evaluate to what extent the negative emotions of service recipients in the healthcare sector may affect negative word of mouth marketing. Based on this purpose, the research model shown in Figure 1 was established.

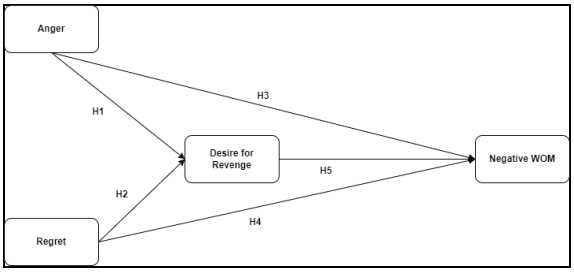


Figure 1. Research Model. Source: Wen-Hai et al. (2019)

Based on the research model, the following hypotheses were determined; H1: Consumer anger has a significant effect on the desire for revenge; H2: Consumer regret has a significant effect on the desire for revenge; H3: Consumer anger has a significant effect on negative WOM; H4: Consumer regret has a significant effect on negative WOM; H5: Consumer desire for revenge has a significant effect on negative WOM. In this study, descriptive survey model, one of the quantitative research methods, was used. A survey was used as the data collection method. The scale developed by Wen-Hai et al. (2019) was used as the survey. The survey form consists of a total of 21 questions: seven questions examining demographic variables, three questions examining the emotion of anger, three questions addressing the feeling of regret, five questions addressing the desire for revenge, and 3 questions addressing negative word-of-mouth. The prepared data collection form first was evaluated by 2 experts in marketing and oral and dental health in terms of content validity, and necessary changes were made in the forms in line with the suggestions. Then, it was applied to a sample group of 60 people via digital form in order to check the understandability and usability of the form. After the preliminary application, the necessary arrangements were made and was given the form its final shape. The population of the research consists of an oral and dental health center operating in Kütahya. A total of 349.116 people have received healthcare services at the center so far. According to the formula used to find the sample number when the research population is known, the sample of this research consists of 381 people. Then, participants were selected through random sampling. The following formula was used for sample calculation; n=(Nt^2 pq)/(d^2 (N-1)+t^2 pq), (N= Size of the total population, n= size of the population sample to obtain, t= Degrees of freedom, p= Frequency of occurrence, q= Frequency of non-occurrence, d= Standard deviation) (Akalpler and Eroğlu, 2015). After obtaining permission from the Kütahya Health Sciences University Non-Interventional Ethics Committee, the survey form was sent to the participants and data was collected. All participants were tried to be reached through both digital and physical survey forms, and a total of 387 people responded. However, five survey forms were excluded from the analysis due to missing information and the analysis continued with a total of 382 surveys.

4. Findings

The data was analyzed with the help of SPSS 21 package program and the following findings were obtained;

Table 1. Demographics of the Respondents

Demographics	Percentage	Demographics	Percentage
Gender		**Famil Income**	
Female	52.60	Less than 17.000 TL	10.20
Male	47.40	17.001-25.000 TL	13.40
Age		25.001-35.000 TL	23.60
18-25 years old	9.20	35.001-45.000 TL	28.00
26-35 years old	19.90	More than 45.001 TL	24.90
36-45 years old	28.50	**Number of Services Received**	
46-55 years old	28.00	1	20.40
56-65 years old	10.70	2	36.60
Over 66 years old	3.70	3	18.30
Education		4 or more	24.60
Literate	2.40	**Kind of Received Procedures**	
Primary School	25.10	Filling Treatment	9.70
High School	32.70	Root Canal Treatment	7.60
Associate Degree	16.00	Implant Treatment	0.80
Bachelor	13.40	Tooth extraction	8.10
Master	8.90	Prosthesis Treatment	4.70
PhD	0.80	Orthodontic Treatment	0.80
Other	0.80	Periodontology	2.60
Marital Status		Teeth Cleaning/Whitening	3.10
Single	30,40	More than a procedure	62.60
Married	69.60		

Looking at the findings in Table 1, it can be seen that the individuals are predominantly married women between the ages of 36-55. When we look at the education levels, it is understood that education is mostly at primary and high school levels. It is seen that individuals with income between 35,001-45,000 TL generally receive more than one service from the hospital and have more than one procedure performed. When the analysis of more than one procedure type was detailed, it was observed that filling, root canal treatment and Gum Disease Treatments (Periodontology) procedures were mostly performed together.

When the reliability levels of the scale and its dimensions are examined, it is seen that the survey scale (0,852) and the sub-dimensions of anger (0,991), desire fr revenge (0,994) and Negative WOM (0,821) have a high alpha value. However, it was determined that the alpha value of the regret (0,146) sub-dimension was extremely low. Since a similar situation occurred in the following analyses, the regret sub-dimension was not included in the analysis.

Table 2. Analysis of Measurement Model

Constructs	**MLE estimates factor loading/measurement error**		**Squared multiple correlation (SMC)**	**Composite reliability (CR)**	**Average of variance extracted (AVE)**
Anger					
A1	0.875	0.234375	0.765625		
A2	0.891	0.206119	0.793881	0.916	0.783
A2	0.889	0.209679	0.790321		
Regret					
R1	0.631	0.601839	0.398161		
R2	0.786	0.382204	0.617796	0.675	0.314
R3	0.488	0.761856	0.238144		
Desire for Revenge					
DR1	0.948	0.101296	0.898704		
DR2	0.956	0.086064	0.913936		
DR3	0.958	0.082236	0.917764	0.980	0.908
DR4	0.956	0.086064	0.913936		
DR5	0.947	0.103191	0.896809		
Negative WOM					
NW1	0.901	0.188199	0.811801		
NW2	0.926	0.142524	0.857476	0.861	0.680
NW3	0.610	0.627900	0.372100		

When we look at the convergent validity analysis in Table 2, which shows the relationships between the expressions of the variables and the factors they form, it is seen that the AVE values of the anger, desire for revenge and negative WOM dimensions are greater than 0.50 and CR values are greater than 0.70. Additionally, AVE values are larger than CR values as expected (Yaşlıoğlu, 2017). However, it was observed that the AVE value of the regret dimension was less than 0.50. For this reason, the regret dimension was not included in the analyzes and the research model was redesigned.

Table 3. Regression Analysis on Hypotheses

Independent	Dependent	Sum. of Model		Anova	Regression Coefficients			H	Result
		R	R2	F	Beta	t	p		
Anger	Desire for Revenge	.786	.618	615.7	.786	24.8	.000	H_1	Accepted
Anger	Negative WOM	.691	.477	347.2	.691	18.6	.000	H_3	Accepted
Desire for Revenge		.816	.666	756.9	0.816	27.5	.000	H_5	Accepted

When we look at the regression analysis in Table 3, desire for revenge dimension is affected by the anger dimension in a significant way (61.80%), the negative WOM dimension is affected by the anger dimension in a significant way (69.10%) and the negative WOM dimension is affected by the desire for revenge dimension in a significant way (81.60%). For this reason, the hypotheses H1, H3 and H5 created at the beginning of the study were accepted. Since both the alpha value and the AVE value of the Regret dimension were very low, they were not included in the analysis. Therefore, hypotheses regarding this dimension were not tested.

5. Conclusion and Suggestions

With this study, it was observed that how negative emotions of individuals affected their negative WOM levels. Accordingly, individuals' anger levels significantly and positively affect both their desire for revenge and their negative WOM levels. Similarly, individuals' desire for revenge significantly and positively affects their negative WOM levels. The results obtained support both the study conducted by Wen-Hai et al. (2019), from which the survey scale was taken, and the studies conducted by Gelbrich (2010), Grégoire et al. (2010), Inman and Zeelenberg (2002). In the light of these findings, it is possible to say that the negative experience of consumers may negatively affect the consumer's attitude towards that company and that he will show a negative WOM tendency by wanting to take revenge on the company during his anger. Therefore, it is important for decision makers both in the health sector and other sectors to take this situation into consideration. Identifying the consumers who has had a negative experience and establishing correct communication with their can prevent the formation of negative word-of-mouth marketing communication about the company.

This study was conducted in an oral and dental health center operating in Kütahya. It is recommended that the study be repeated in health institutions providing different services. Again, since the city where the study was conducted is small and conservative, it may be recommended to repeat a similar study in larger and metropolitan cities. Thus, it can be observed whether there is a change according to culture. Except those, in the study, only the concepts of anger and desire for revenge, among the negative emotions, were applied. The study can be

expanded by including different emotions such as discontent, dislike, embarrassment, sadness, and worry (Romani et al., 2012).

Considering the demographic characteristics of the participants in the study, it is seen that they generally consist of people with low income levels and who receive free healthcare services. It is recommended to repeat the same study from a private hospital and compare the differences.

References

Akalpler, Ö., & Eroğlu, K. (2015). Kuzey Kıbrıs Türk Cumhuriyeti'nde üniversite öğrencilerinin sık görülen cinsel yolla bulaşan enfeksiyonlara ilişkin bilgileri ve cinsel davranışları. Hacettepe Üniversitesi Hemşirelik Fakültesi Dergisi, 2(2), 1-19.

Centers for Disease Control and Prevention. (2011). What is health marketing? Accessed November 13, 2014, available from http://www.cdc.gov/healthcommunication/toolstemplates/whatishm.html. Cited in Swenson, E. R., Bastian, N. D., & Nembhard, H. B. (2018). Healthcare market segmentation and data mining: A systematic review. Health marketing quarterly, 35(3), 186-208.

Chaniotakis, I. E. and Lymperopoulos, C. (2009). Service quality effect on satisfaction and word of mouth in the health care industry. Managing Service Quality, 19(2), 229-242.

Curwen, L. G., & Park, J. (2014). When the shoe doesn't fit: female consumers' negative emotions. Journal of Fashion Marketing and Management, 18(3), 338-356.

Gelbrich, K. (2010). Anger, frustration, and helplessness after service failure: coping strategies and effective informational support. Journal of the Academy of Marketing Science, 38, 567-585.

Greenspun, H., & Coughlin, S. (2012). The U.S. health care market: a strategic view onconsumer segmentation. Deloitte Center for Health Solutions. Accessed November 20, 2015, available fromhttp://www2.deloitte.com/content/dam/Deloitte/us/Documents/life-sciences-health-care/us-lhsc-mhealth-in-an-mworld-103014.pdf. Group Press, Austin, TX.

Grégoire, Y., Laufer, D., & Tripp, T. M. (2010). A comprehensive model of customer direct and indirect revenge: Understanding the effects of perceived greed and customer power. Journal of the Academy of Marketing Science, 38, 738-758.

Helena Vinagre, M., & Neves, J. (2008). The influence of service quality and patients' emotions on satisfaction. International journal of health care quality assurance, 21(1), 87-103.

Inman, J. J., & Zeelenberg, M. (2002). Regret in repeat purchase versus switching decisions: The attenuating role of decision justifiability. Journal of consumer research, 29(1), 116-128.

Kay, M. J. (2007). Healthcare marketing: what is salient?. International Journal of Pharmaceutical and Healthcare Marketing, 1(3), 247-263.

Khalid, S., Ahmed, M. A., & Ahmad, Z. (2013). Word-of-mouth communications: A powerful contributor to consumers decision-making in healthcare market. International journal of business and management invention, 2(5), 55-64.

Litvin, S.W., Goldsmith, R.E. & Pan, B. (2008). Electronic word-of-mouth in hospitality and tourism management. Journal of Tourism Management, 29(3), 458-68.

Martin, S. (2017). Word-of-mouth in the health care sector: a literature analysis of the current state of research and future perspectives. International Review on Public and Nonprofit Marketing, 14, 35-56.

Pauli, G., Martin, S., & Greiling, D. (2023). The current state of research of word-of-mouth in the health care sector. International Review on Public and Nonprofit Marketing, 20(1), 125-148.

Reinhardt, U. E. (2005). Variations In California hospital regions: Another wake-up call for sleeping policymakers: How long must John Wennberg and colleagues howl in the wind before the policy community pays attention to their critical findings?. Health Affairs, 24(Suppl1), W5-549.

Richins, M. L. (1997). Measuring emotions in the consumption experience. Journal of consumer research, 24(2), 127-146.

Romani, S., Grappi, S., & Dalli, D. (2012). Emotions that drive consumers away from brands: Measuring negative emotions toward brands and their behavioral effects. International Journal of Research in marketing, 29(1), 55-67.

Swenson, E. R., Bastian, N. D., & Nembhard, H. B. (2018). Healthcare market segmentation and data mining: A systematic review. Health marketing quarterly, 35(3), 186-208.

Trusov, M., Bucklin, R. E., & Pauwels, K. (2009). Effects of word-of-mouth versus traditional marketing: findings from an internet social networking site. Journal of marketing, 73(5), 90-102.

Wen-Hai, C., Yuan, C. Y., Liu, M. T., & Fang, J. F. (2019). The effects of outward and inward negative emotions on consumers' desire for revenge and negative word of mouth. Online Information Review, 43(5), 818-841.

Whyte, J.W.H. (1954). The web of word-of-mouth. Fortun, 48-112.

Woodside, A., Nielson, R., Walters, R., & Muller, G. (1998). Preference segmentation ofhealth care services: the old-fashioneds, value conscious, affluents, and professionalwant-it-alls.Journal of Health Care Marketing,8(2), 14–24.

Yaşlıoğlu, M. M. (2017). Sosyal bilimlerde faktör analizi ve geçerlilik: Keşfedici ve doğrulayıcı faktör analizlerinin kullanılması. İstanbul Üniversitesi İşletme Fakültesi Dergisi, 46, 74-85.

Yi, S., & Baumgartner, H. (2004). Coping with negative emotions in purchase-related situations. Journal of Consumer psychology, 14(3), 303-317.

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17796

Improving Accuracy in Geospatial Information Transfer: A Population Density-Based Approach

Virgilio Pérez, Jose M. Pavía
Applied Economics, University of Valencia, Spain.

How to cite: Pérez, V.; Pavía J.M. 2024. Improving Accuracy in Geospatial Information Transfer: A Population Density-Based Approach. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17796

Abstract

The R package sc2sc offers fundamental tools for transferring information between census sections and postal codes in Spain, based on the cartography of these geographic segmentations. However, certain aspects for improvement have been identified. This document presents a substantial improvement to the package, optimizing the cp2sc function, which facilitates the transfer of information from postal codes to census sections. The introduced improvement considers population density as a corrective factor in the process, resulting in a more accurate and relevant data allocation. Various use cases highlight the improvement of the new methodology, though they also underline the need to work with updated and precise cartography, opening new lines of research and future work.

Keywords: *Spatial Statistics; Longitudinal Data; Census Sections; Postal Codes; Sc2sc; R-Stats; Geospatial Analysis.*

1. Introduction

In Spain, several geographic segmentations coexist. Census sections represent the lowest level of spatial aggregation used by the National Statistics Institute (INE). At this level, the INE collects and distributes demographic and socioeconomic data, including statistics related to population and housing censuses. These geographical units, which generally group between 1,000 and 2,000 people, are crucial for the detailed study of population characteristics, allowing precise segmentation for urban planning and public policy development (Pillet et al., 2013). On the other hand, postal codes, assigned by Correos (the national postal service of Spain), are designed to optimize the postal distribution system and can cover areas ranging from small urban zones to larger regions in non-urban areas (Otero et al., 2019). Postal codes, while considering the spatial distribution of the population across the territory, reflect the organization of the postal network.

Although both geographical systems serve different purposes, they are indispensable for various applications (such as logistics, marketing, and socioeconomic analysis, among others), offering complementary perspectives on the distribution and organization of the population and services in the country (Reina-Usuga et al., 2020). The coexistence of both geographic segmentation systems implies that certain information is available for one level of aggregation but not for the other (Manzanares & Riquelme, 2017). This complicates, if not directly prevents, analysis that includes certain variables, hindering the work of researchers and other social agents.

In this context, having a tool that allows the transfer of information over time between census sections and/or postal codes could transform the way researchers and analysts can understand population and territorial dynamics. Such capability enables a more detailed and accurate analysis of demographic, economic, and social trends over time, facilitating decision-making and the formulation of public policies. For example, by analyzing how the population characteristics of a specific area evolve, urban planners can design more inclusive and sustainable cities, while economists can assess the impact of economic policies in different regions (Ropero et al., 2015; Thompson et al., 2020).

Aiming to offer this analytical capability, the authors of the present document have developed the sc2sc R package (Pérez & Pavía, 2024), a tool that facilitates the transfer of information between census sections and postal codes in Spain, using the cartography (territorial boundaries) of both geographic segmentation methods. This package exports three functions: i) sc2sc, that allows the transfer of statistics between census sections that coincide (totally or partially) at different points in time; ii) sc2cp, that enables the transfer of information from census sections to postal codes; and iii) cp2sc, that facilitates the transfer of information from postal codes to census sections.

The functionalities offered by sc2sc are based on an approach that integrates and analyzes the cartography of census sections, provided by the INE (2024a), and the cartography of postal codes generated by Goerlich (2022). The basic solutions it offers are based on the intersections between polygons representing census sections in consecutive years, or between the polygons of census sections and postal codes from 2019. This approach allows for the transfer of statistics between census sections from different time periods, as well as between census sections and postal codes and vice versa, covering a time range from 2001 to 2023, excluding 2002.

However, we have observed that when transferring information from postal codes to census sections using the basic solution, the distribution is not consistent. Working under the assumption that statistics (variables) are uniformly distributed throughout the territory of a postal code, we are assigning greater importance (weight) to those census sections that are larger, when it is common that the census sections with greater extension are those with less population density.

This document provides a possible solution to this problem by including population density in the mathematical formula that allows for the transfer/imputation of statistics to census sections from postal codes. The rest of the document is structured as follows: In the second section, we present the implemented methodology. In the third section, we demonstrate the opportunity of the new approach by developing some examples. In the fourth section, the work is discussed, and brief conclusions and some ideas for future work are presented.

2. Methodology

A postal code (cp), spatially defined as a polygon, intersects with $n \geq 1$ census sections (sc), also polygons. Thus, if we want to transfer/distribute/impute a value L available in cp among sc, we could leverage in the relationship

$$L = \frac{|sc_1 \cap cp|}{|cp|} \cdot L + \frac{|sc_2 \cap cp|}{|cp|} \cdot L + \cdots + \frac{|sc_n \cap cp|}{|cp|} \cdot L, \quad (1)$$

meeting the condition

$$\frac{|sc_1 \cap cp|}{|cp|} + \frac{|sc_2 \cap cp|}{|cp|} + \cdots + \frac{|sc_n \cap cp|}{|cp|} = 1, \quad (2)$$

where $|A|$ denotes the area of polygon A, to make the transfer.

However, as mentioned in the introduction, this method is not consistent in terms of reversibility for much of the Spanish territory. This is due to the different criteria implemented to configure each territorial segmentation. While census sections group between 1,000 and 2,000 inhabitants, the delimitation of postal codes presents more heterogeneity in terms of resident. To address this imbalance, a significant improvement has been introduced in the methodology: weighting by population density. This adjusts the allocation of information not only based on the area of the census section but also considering its population density.

Let $D_i = P_i/A_i$ be the population density of the target census section, obtained as the quotient between the resident population in that census section, P_i, and the area, A_i, of the polygon. To transfer a value L from cp to sc we can take into account D_i, as it is verified that:

$$L = D_1 \cdot \frac{|sc_1 \cap cp|}{P_{cp}} \cdot L + D_2 \cdot \frac{|sc_2 \cap cp|}{P_{cp}} \cdot L + \cdots + D_n \cdot \frac{|sc_n \cap cp|}{P_{cp}} \cdot L, \quad (3)$$

where $P_{cp} = \sum_{i=1}^{n} D_i \cdot |sc_i \cap cp|$. Meeting the condition that:

$$D_1 \cdot \frac{|sc_1 \cap cp|}{P_{cp}} + D_2 \cdot \frac{|sc_2 \cap cp|}{P_{cp}} + \cdots + D_n \cdot \frac{|sc_n \cap cp|}{P_{cp}} = 1. \quad (4)$$

Levearing on this expressions allows for a more equitable and representative distribution of information, reflecting more accurately the real distribution of the population and its characteristics within the geographic units of analysis.

3. Results

To assess the methodology described in the previous section, several examples have been developed, checking the consistency of the proposed approach based on the number of inhabitants per census section in 2019. Figure 1 depicts the municipality of Alaquàs (Valencia), with postal code 46970 and municipality code 46005. The blue line indicates the boundary of the postal code, while the red lines reflects the boundaries of the 21 census sections that make up the municipality. In this case, the correspondence between postal codes and census sections is 1:n, meaning i) a single postal code encompasses all the census sections of the municipality, and ii) the census sections that make up the represented postal code do not intersect with any other postal code. As can be observed, the extension of the census sections is very unequal, with the section coded 02001 occupying the most territory of the municipality.

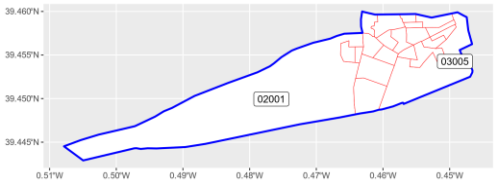


Figure 1. Territorial boundaries (census sections and postal code) of Alaquàs (Valencia) in 2019. Source: own elaboration based on the cartography from INE (2024b) and Goerlich (2022).

In this example, given the structure and composition of the involved polygons, we can assign to postal code 46970 a population equivalent to the sum of the populations of the 21 previously mentioned census sections. With this, we proceed to transfer this information, at the postal code level, to the relevant census sections, using the cp2sc function from the R package sc2sc using version 12, which is based on areal weighting, and using the new latest version, which includes the improvement.

In Table 1, on one hand, the official population of each census section in 2019 is provided (variable POP), as well as the area occupied by each portion of the territory and the ratio to the total (variables AREA and RATIO); and on the other hand, the population estimates calculated, both with the previous method (version 12) and with the method proposed in this document (variables POP_prev and POP_new) alog with the differences, in absolute terms, from the real values (variables diff_prev and diff_new). As can be observed, the previous method, based solely on the proportion of territory occupied by each census section in relation to the extension of the postal code, reports inconsistent values. In contrast, and as would be expected, the new method proposed returns exactly the same population from which it started.

Table 1. Population transfer, by census sections, of the municipality of Alaquàs (Valencia) in 2019.
Source: own elaboration based on the continuous register statistics (INE, 2024b).

CUSEC	POP	AREA	RATIO	POP_prev	diff_prev	POP_new	diff_new
4600501001	1221	0.05	0.012	361	860	1221	0
4600501002	1068	0.08	0.020	588	480	1067	1
4600501003	1349	0.05	0.013	387	962	1348	1
4600501004	1103	0.05	0.013	393	710	1103	0
4600501005	1148	0.04	0.009	278	870	1147	1
4600501006	1238	0.03	0.007	210	1028	1239	1
4600501007	715	0.02	0.004	127	588	715	0
4600501008	723	0.02	0.005	151	572	724	1
4600501010	1934	0.06	0.015	435	1499	1933	1
4600502001	2133	2.34	0.601	17757	15624	2131	2
4600502002	1743	0.06	0.014	423	1320	1744	1
4600502003	1392	0.04	0.010	304	1088	1392	0
4600502004	2515	0.12	0.032	934	1581	2516	1
4600502005	1667	0.05	0.012	352	1315	1667	0
4600502006	1002	0.02	0.004	133	869	1002	0
4600502007	1308	0.15	0.038	1123	185	1307	1
4600503001	1022	0.03	0.007	213	809	1023	1
4600503002	1243	0.03	0.006	192	1051	1242	1
4600503003	1248	0.03	0.008	225	1023	1247	1
4600503004	1701	0.08	0.020	600	1101	1700	1
4600503005	2088	0.58	0.148	4372	2284	2087	1
TOTAL/ERROR	**29561**	**3.90**	**1.000**	**29558**	**35819**	**29555**	**16**

A second example has been developed following the procedure outlined for the previous case. In this instance, the municipality of Alboraia (Valencia), with postal code 46120 and municipality code 46013, was taken as a reference. As can be seen in Figure 2, we again find differences between the two cartographies, primarily in the eastern boundaries of the municipality (beach/coast), generating certain inconsistencies in the results.

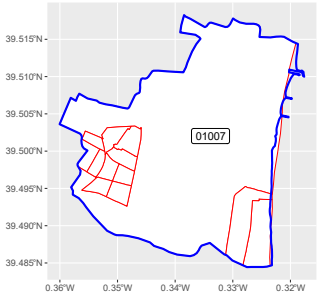


Figure 2. Territorial boundaries (census sections and postal code) of Alboraia (Valencia) in 2019. Source: own elaboration based on the cartography from INE (2024b) and Goerlich (2022).

Table 2. Transfer ot number of voters, residents in Alboraia (Valencia), who participated in the 2019 Valencian Courts elections. Source: own elaboration based on the results of the 2019 Valencian Courts elections (Pérez et al., 2021).

CUSEC	VOTES	AREA	RATIO	VOT_prev	diff_prev	VOT_new	diff_new
4601301001	837	0.09	0.011	152	685	811	26
4601301002	690	0.11	0.013	185	505	662	28
4601301003	908	0.09	0.011	151	757	890	18
4601301004	1509	0.07	0.009	123	1386	1407	102
4601301005	975	0.05	0.006	82	893	1050	75
4601301006	999	0.08	0.009	131	868	1013	14
4601301007	1250	6.89	0.827	11668	10418	1365	115
4601301008	598	0.07	0.008	120	478	557	41
4601301009	937	0.06	0.007	100	837	978	41
4601301010	1565	0.07	0.008	114	1451	1482	83
4601301011	1204	0.09	0.010	150	1054	1120	84
4601301012	1330	0.33	0.040	548	782	1382	52
4601301014	1310	0.35	0.042	588	722	1393	83
4625011033			0.000	0	0	0	0
4625011039			0.000	0	0	0	0
4625011050			0.000	0	0	0	0
TOTAL/ERROR	**14112**	**8.33**	**1.000**	**14112**	**20836**	**14110**	**762**

In this example, we illustrate the differences using a variable of a different kind. In particular, we consider the number of voters who participated in the 2019 Valencian Courts elections. This

case notably elucidates the versatility of sc2sc in addressing analysis in different fields, including the political-electoral sphere.

From these data and the results obtained using the cp2sc function, Table 2 was constructed. Again, the values obtained after applying the methodology proposed in this paper fit more closely with reality (the transfer error is this time more than 27 times smaller). Also in this case, as in the previous one, differences between the estimated values and the actual values (number of voters) are observed, primarily derived from divergences in turnout rates. In addition to the above, and due to cartography issues, values were imputed to three census sections of neighboring municipalities, which, although it did not imply any significant imputation, highlights the relevance of working with consistent geographical data.

4. Discussion and Conclusions

The implementation of the sc2sc package has highlighted the complexity and challenges inherent in the task of transferring information between different levels of geographic aggregation, especially between census sections and postal codes in Spain. The solution proposed in this document, which incorporates population density into the process of transferring information from postal codes to census sections, not only addresses a key limitation of an approach based exclusively on surfaces but also highlights the gains in accuracy that can be obtained in the transferred data, which align more closely with demographic and geographic reality.

The examples presented illustrate the improvements of the proposed methodology in different municipal contexts, demonstrating its ability to more faithfully reflect the distribution of the population within the census sections, even in scenarios of complex correspondence between postal codes and census sections. However, the differences observed between the estimated and actual populations underscore the critical importance of having precise and updated cartographies, as well as the need for additional adjustments to improve the congruence between the geographical data used in the analysis.

Future work should focus on refining these aspects, exploring advanced techniques for handling cartographic discrepancies, and optimizing information transfer algorithms to increase their applicability and accuracy. Likewise, expanding the temporal range covered and including new variables could significantly enrich the analytical capabilities of the sc2sc package, opening new avenues for research and geospatial analysis in Spain.

References

Goerlich, F. (2022). Elaboración de un mapa de Códigos Postales de España con recursos libres: cómo evitar pagar por disponer de información de referencia. IVIE working papers, 2022:03. doi.org/10.12842/WPIVIE_0322

INE (2024a). Instituto Nacional de Estadística. Cartografía secciones censales y callejero de Censo Electoral. Retrieved from: https://www.ine.es/uc/1dIJtjmE

INE (2024b). Instituto Nacional de Estadística. Estadística del Padrón continuo. Retrieved from: https://www.ine.es/uc/XUXtgIDdi1

Manzanares, A. and Riquelme, P.J. (2017). Análise espacial do desemprego nos mercados locais de traballo españois. *Revista galega de economía: Publicación Interdisciplinar da Facultade de Ciencias Económicas e Empresariais*. *26*(2), 29-42.

Otero, R., García, J., Domínguez, J., and Pérez, A. (2019). Inmigración y dinámicas territoriales en España: crisis y recuperación (2008-2017). In E. Mañé (Ed.), *Anuario CIDOB de la Inmigración (2019)* (pp. 190-217). Barcelona: Ediciones Bellaterra. doi.org/10.24241/AnuarioCIDOBInmi.2019.190

Pérez, V. and Pavía, J.M. (2024). sc2sc: Spatial Transfer of Statistics among Spanish Census Sections. Version 0.0.1-12. https://cran.r-project.org/package=sc2sc

Pérez, V., Aybar, C., and Pavía, J.M. (2021). Spanish electoral archive. SEA database. *Scientific Data*, *8*, 193. doi.org/10.1038/s41597-021-00975-y

Pillet, F., Cañizares, M.C., Ruiz, A.R., Martínez, H.S., Plaza, J., and Santos, J.F. (2013). Los indicadores de la cohesión territorial en el análisis de la escala supramunicipal o subregional. *Ería: Revista cuatrimestral de geografía*, *90*, 91-106.

Reina-Usuga, L., Haro-Giménez, T., and Parra-López, C. (2020). Food governance in Territorial Short Food Supply Chains: Different narratives and strategies from Colombia and Spain. *Journal of Rural Studies*, *75*, 237-247. doi.org/10.1016/j.jrurstud.2020.02.005

Ropero, R.F., Aguilera, P.A., and Rumí, R. (2015). Analysis of the socioecological structure and dynamics of the territory using a hybrid Bayesian network classifier. Ecological Modelling, 311, 73-87. doi.org/10.1016/j.ecolmodel.2015.05.008

Thompson, M., Nowak, V., Southern, A., Davies, J., and Furmedge, P. (2020). Re-grounding the city with Polanyi: From urban entrepreneurialism to entrepreneurial municipalism. *Environment and Planning A: Economy and Space*, *52*(6), 1171-1194. doi.org/10.1177/0308518X19899698

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17462

Bibliometrics and Scientometrics of the Business Agility

Petra Lesníková, Andrea Janáková Sujová
Department of Economics, Management and Business, Technical University in Zvolen, Slovakia

How to cite: Lesníková, P.; Janáková Sujová A. 2024. Bibliometrics and Scientometrics of the Business Agility. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17462

Abstract

Bibliometric analysis is an important tool in scientific research designed to explore and analyse a range of scientific data. The purpose of this paper is to highlight the growing importance and relevance of business agility issues in the scientific community. The aim is to provide a brief insight into business agility through bibliometric analysis of articles included in the WOS and Scopus databases. As a result, a comparison of these databases is presented along with a description of the resulting clusters using the software tools VOSviewer and SciMAT. The area of interest in the databases is Business, Economics, Management and Finance in the publication years 1994-2023. The results show that although the databases overlap to some extent, there are some slight differences in terms of bibliometrics or scientometrics. Although the Scopus database had a higher number of publications, the number of keyword occurrences is higher in the database WOS. There are also slight differences in the most numerous keywords. In terms of clusters, the number is the same, but slight differences are also observed. Based on the analysis of the occurrence of keywords, it is possible to note an increased interest in the issue of agility, which is linked to a number of other areas of management. The Scopus database is recommended to study business agility.

Keywords: *Bibliometric Analysis; Scientometrics; Business Agility; Keywords; Clusters.*

1. Introduction

Bibliometric analysis is an important tool in scientific research designed to explore and analyze large amounts of scientific data, including developments and trends in certain fields (Donthu et al., 2021). At the same time, it is used in various contexts ranging from information science, medicine and business and management. From the current perspective, the two main subjects of bibliometric interest are Information Science & Library Science and Computer Science (Lyu et al., 2023). According to Godin (2006, p. 109), "Among the many statistics on science, called scientometrics, bibliometrics holds a privileged place." Descriptive bibliometrics mainly focuses on tracking the number of articles (for possible comparisons); evaluative bibliometrics looks at how articles have influenced the subsequent research of others (McBurney & Novak, 2002). Science mapping is an analytical technique that is effective at mapping the strength of associations between information items, while it allows to highlight potentially significant patterns or trends of scientific change that can guide the exploration and interpretation of visualised structures (Chen, 2017). Bibliometric analysis can make use of a rather broad methodological apparatus, which has been clearly elaborated by Donthu et al. (2021). The primary techniques of bibliographic analysis include performance analysis and scholarly mapping, which include common word analysis, co-authorship analysis, citation analysis, co-citation analysis, or bibliographic linkage. Software tools for performing bibliometric analysis of scientific mapping include BibExcel, Biblioshiny, BiblioMaps, CiteSpace, CitNetExplorer, SciMAT, Sci^2 Tool, and VOSviewer (Moral-Muñoz et al., 2020).

The term scientometrics is often used in connection with bibliometric analysis. Scientometrics is commonly referred to synonymously in the literature as bibliometrics, infometrics or scientific mapping. According to Yang & Yuan (2017), bibliometric, scientometric and infometric "differ in the degrees of utilization and recognition but are similar in the general direction". Examples of publication metrics used to assess scholarly productivity, impact and relevance are Impact factor, h-index, Journal impact quartile, Article Influence Score, CiteScore. Publication metrics are not only important for the careers of individuals, but also influence the progress of science as a whole through their role in the award process (Myers & Kahn, 2021).

Bibliometric and scientometric are also applicable in the fields of Management, Business and Economics. There are concepts that are gaining importance in the field at different times. This also applies to the concept of business agility. In terms of the essence of the concept, business agility is understood as an organization's capacity to adapt quickly to changing market dynamics, customer demands, and industry standards profitably and cost-effectively without compromising on quality (Yusuf et al., 2023). Business agility is more than just flexibility or adaptability. According to Van Oosterhout et al. (2006) it is the ability to quickly and easily change businesses and business processes beyond the normal level of flexibility to effectively

manage unpredictable external and internal changes. The Business Agility Institute (2023), based on a survey of businesses, listed the most significant organisational benefits of business agility, namely business outcomes and value, customer satisfaction, adaptability to change, employee satisfaction, process improvements and others.

2. Methodology

Bibliometrics and scientometrics have a wide range of potential applications in various fields. The aim of this article is to provide a brief insight into business agility through bibliometric analysis of articles included in the WOS and Scopus databases. As a result, a comparison of these databases is presented along with a brief description of the resulting clusters using the software tools VOSviewer and SciMAT. The area of interest in the WOS database was Business & Economics; Management; Finance (WOS categories), Business agility (topic, keyword), years 1994-2023 (publication years). A similar procedure was followed for the Scopus database: Business agility (article title, abstract, keywords), Business, Management and Accounting; Economics, Econometrics and Finance (subject area) and years 1993-2023 (publication years).

After obtaining the datasets, a results analysis (types of documents, publication years, countries/regions) and performance analysis (number of documents, number of citations) was performed. Science mapping was performed using VOSviewer. The map was created using co-occurrence analysis of keywords and the full count method. A minimum number of keyword occurrences at level of 10 was chosen. According to Van Eck & Waltman (2022), the attributes Links and Total link strength indicate the number of links of an article with other articles and the total strength of the links of an article with other articles. The result is a cluster map that can be analyzed from multiple perspectives (in our case only network visualization is used). The minimum cluster size was chosen to be 10 items. Subsequently, the SciMAT program was used, which represents an open-source software tool, to perform longitudinal scientific mapping developed by Cobo et al. (2012). Through this program, it was used only a partial analysis, of keywords occurrence. The conclusion briefly discusses publication metrics and a comparison of relevant databases.

3. Results

3.1 Brief view of business agility from publications of WOS database

After document selection (Methodology section), the sample consisted of 1,283 publications from the Web of Science Core Collection. In terms of WOS categories, the most represented areas are Management (823; 64.15%) and Business (314; 24.47%). The total number of publications includes 867 articles (67.58%) and 372 proceedings papers (29%). Slight fluctuations are shown over the years. Basically, there is an increase in publications in four

waves (2002-2005), (2006-2009), (2012-2015) and (2018-2023), with the highest number of publications in 2023 (139). The largest number of publications are from USA (267; 20.81%), People R China (155;12.08%) and England (147; 11.46%).

Based on keywords occurrence analysis, 4,082 keywords were recorded, of which 175 meet the threshold (minimum number of occurrences is 10). The network visualization of these keywords is shown in Figure 1.

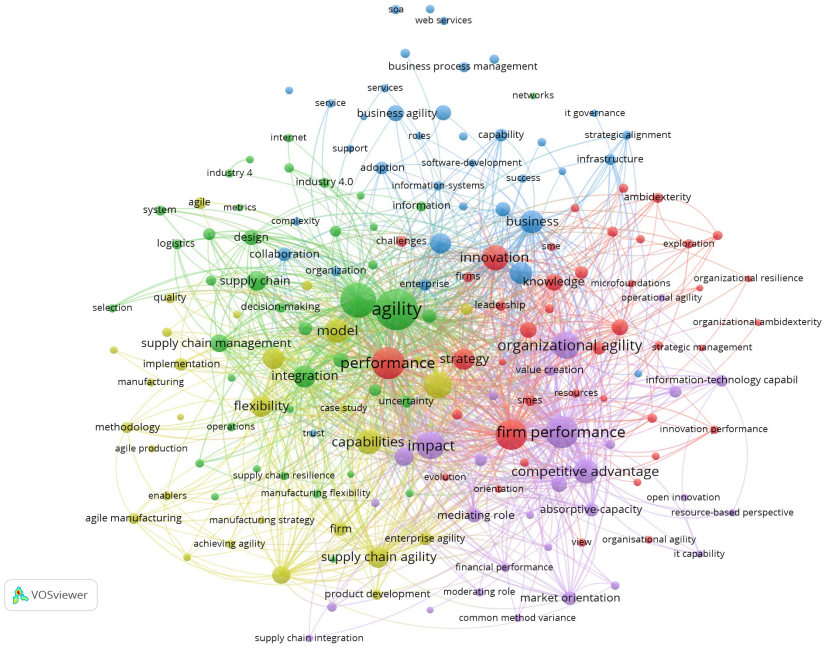


Figure 1. Network visualization of keywords in topic of business agility (WOS). Source: own processing.

From the figure, the keyword with the highest frequency is clearly the word agility (319 occurrences; 173 links, 1,971 total link strenght). The 175 items are divided into 5 clusters. Cluster 1 (red colour; 41 items) is composed of keywords mostly strategically oriented. It includes concepts such as dynamic capabilities, performance, strategy, strategic agility, innovation. Cluster 2 (green colour; 40 items) focuses mainly on agility in the general sense in connection with industry (also industry 4.0) and supply chain management. Cluster 3 (blue colour; 36 items) focuses on business agility, processes, infrastructure and technology. Cluster 4 (yellow colour; 29 items) includes concepts oriented towards agile manufacturing and production. In this context, business performance and IT are also linked. Cluster 5 (purple colour; 29 items) mainly focuses on organisational agility, operational agility in the context of firm performance and resource based view. This shows that there are slight differences between the clusters, while it is just interesting to look at the existing word linkages between the clusters.

Business agility in particular has links with the keywords performance, organizational agility, IT, and capabilities.

3.2 Brief view of business agility from Scopus database

The sample of publications from the Scopus database consisted of 1,586 publications. The most represented subject area (similar to WOS categories) is Business, Management, Accounting (1,532). The total number of publications includes 951 articles (59.96%) and 271 conference papers (17.09%). More frequent fluctuations are shown over the years studied than in the case of the WOS database, but with the exception of the fluctuations mentioned above, there is still a gradual increase in publications. There is also a slight change from the WOS database in the countries from where the increase in publications comes from, namely USA (297; 18.73%), United Kingdom (164; 10.34%) and India (143; 9.02%). China is reported in only 80 cases (5.04%). From the perspective of keywords occurrence analysis, of 5,860 keywords, 135 meet the threshold. As in the case of the WOS database, 5 clusters emerged (Figure 2).

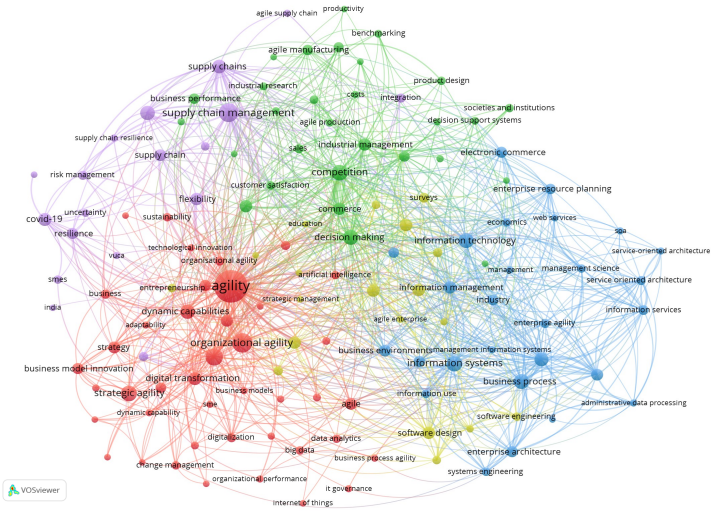


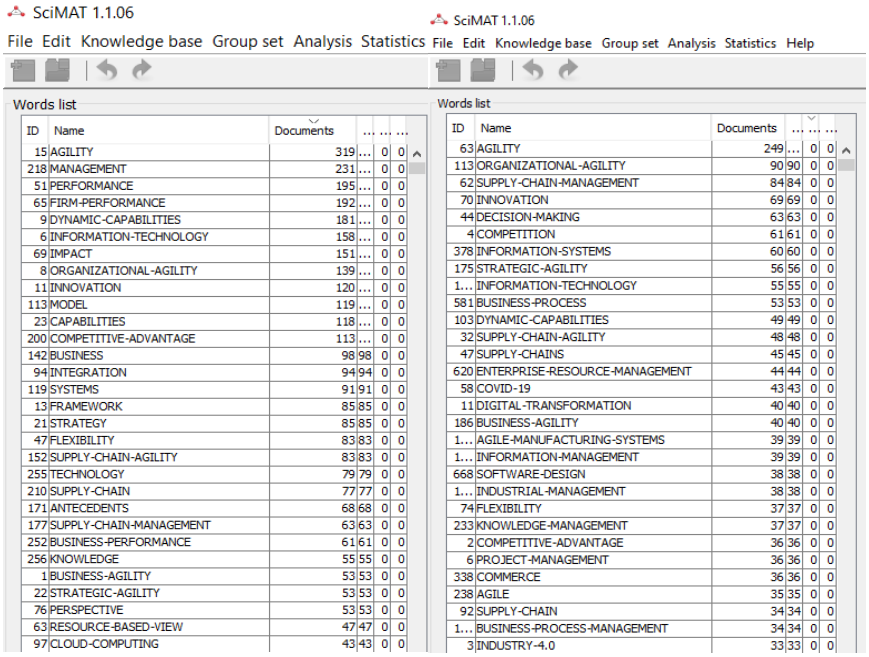
Figure 2. Network visualization of keywords in topic of business agility (Scopus). Source: own processing.

The keyword with the highest frequency is the word agility (249 occurrences; 125 links, 575 total link strenght). Cluster 1 (red colour; 42 items) is composed of keywords mostly of a broader and more general nature. It includes terms such as agility, organisational agility, strategic agility, performance, business model, or dynamic capabilities. Cluster 2 (green colour; 29 items) focuses mainly on manufacturing and industry (industrial management, agile manufacturing, etc.). Cluster 3 (blue colour; 27 items) focuses on the systems aspect, in the form of infrastructure systems, IT, systems engineering, etc. Cluster 4 (yellow colour; 19 items) is close to the previous cluster and includes agile enterprise concepts related to software. These are in

particular software design, software engineering and process management. The cluster also includes business agility with 39 occurrences, and has more links than the previous one. Cluster 5 (purple colour; 18 items) focuses on supply chain management (other keywords are e.g. flexibility, resilience, covid-19).

3.3 Comparison of WOS and Scopus database results

To complete the keyword occurrence analysis, Figure 3 shows a comparative overview of the 30 keywords with the highest number of occurrences from both databases. The given overview is generated by SciMAT, and the keywords were kept original, i.e., we did not group them.


SciMAT 1.1.06

File Edit Knowledge base Group set Analysis Statistics

Words list

ID	Name	Documents	...	...	...
15	AGILITY	319	...	0	0
218	MANAGEMENT	231	...	0	0
51	PERFORMANCE	195	...	0	0
65	FIRM-PERFORMANCE	192	...	0	0
9	DYNAMIC-CAPABILITIES	181	...	0	0
6	INFORMATION-TECHNOLOGY	158	...	0	0
69	IMPACT	151	...	0	0
8	ORGANIZATIONAL-AGILITY	139	...	0	0
11	INNOVATION	120	...	0	0
113	MODEL	119	...	0	0
23	CAPABILITIES	118	...	0	0
200	COMPETITIVE-ADVANTAGE	113	...	0	0
142	BUSINESS	98	98	0	0
94	INTEGRATION	94	94	0	0
119	SYSTEMS	91	91	0	0
13	FRAMEWORK	85	85	0	0
21	STRATEGY	85	85	0	0
47	FLEXIBILITY	83	83	0	0
152	SUPPLY-CHAIN-AGILITY	83	83	0	0
255	TECHNOLOGY	79	79	0	0
210	SUPPLY-CHAIN	77	77	0	0
171	ANTECEDENTS	68	68	0	0
177	SUPPLY-CHAIN-MANAGEMENT	63	63	0	0
252	BUSINESS-PERFORMANCE	61	61	0	0
256	KNOWLEDGE	55	55	0	0
1	BUSINESS-AGILITY	53	53	0	0
22	STRATEGIC-AGILITY	53	53	0	0
76	PERSPECTIVE	53	53	0	0
63	RESOURCE-BASED-VIEW	47	47	0	0
97	CLOUD-COMPUTING	43	43	0	0

SciMAT 1.1.06

File Edit Knowledge base Group set Analysis Statistics Help

Words list

ID	Name	Documents	...	...	...
63	AGILITY	249	...	0	0
113	ORGANIZATIONAL-AGILITY	90	90	0	0
62	SUPPLY-CHAIN-MANAGEMENT	84	84	0	0
70	INNOVATION	69	69	0	0
44	DECISION-MAKING	63	63	0	0
4	COMPETITION	61	61	0	0
378	INFORMATION-SYSTEMS	60	60	0	0
175	STRATEGIC-AGILITY	56	56	0	0
1...	INFORMATION-TECHNOLOGY	55	55	0	0
581	BUSINESS-PROCESS	53	53	0	0
103	DYNAMIC-CAPABILITIES	49	49	0	0
32	SUPPLY-CHAIN-AGILITY	48	48	0	0
47	SUPPLY-CHAINS	45	45	0	0
620	ENTERPRISE-RESOURCE-MANAGEMENT	44	44	0	0
58	COVID-19	43	43	0	0
11	DIGITAL-TRANSFORMATION	40	40	0	0
186	BUSINESS-AGILITY	40	40	0	0
1...	AGILE-MANUFACTURING-SYSTEMS	39	39	0	0
1...	INFORMATION-MANAGEMENT	39	39	0	0
668	SOFTWARE-DESIGN	38	38	0	0
1...	INDUSTRIAL-MANAGEMENT	38	38	0	0
74	FLEXIBILITY	37	37	0	0
233	KNOWLEDGE-MANAGEMENT	37	37	0	0
2	COMPETITIVE-ADVANTAGE	36	36	0	0
6	PROJECT-MANAGEMENT	36	36	0	0
338	COMMERCE	36	36	0	0
238	AGILE	35	35	0	0
92	SUPPLY-CHAIN	34	34	0	0
1...	BUSINESS-PROCESS-MANAGEMENT	34	34	0	0
3	INDUSTRY-4.0	33	33	0	0

Figure 3. Top 30 keywords of business agility from WOS database (left) and Scopus database (right). Source: own processing

The keyword comparison shows that despite the fact that fewer publications remained from the WOS database after filtering (1,283) compared to the Scopus database (1,586) - the same applies to the number of keywords (4,082 keywords in the case of WOS, 5,860 keywords in the case of Scopus) - the number of occurrences is higher in the case of WOS . The number of occurrences for the word agility is 319. The nature of the first keywords in this case is rather more general - management, performance, dynamic capabilities, information technology. It is only on the lower rungs that terms such as organizational agility or strategic agility appear. In the case of keywords from the Scopus database, the situation is different. The word agility appears in 249 documents and the second most frequently mentioned word has only 90 occurrences. Another difference is

that the keywords are already rather more specific (organizational agility, supply chain agility, etc.), while keywords such as (financial) performance occur 30 times, management (science) has only 18 occurrences (not shown in the figure). A comparison of the obtained dataset results according to the selected characteristics is presented in Table 1.

Table 1. Comparison of datasets in bibliometric analysis. Source: own processing

Characteristics	WOS database	Scopus database
Number of publications	1,283	1,586
Sum of times cited	37,633	40,146
Average citations per item	29.33	25.31
H-index	93	93
Fluctuations in the number of publications/years	slightly	more often
Countries with the highest number of publications	USA, China, UK	USA, UK, India
Number of total keywords	4,082	5,860
Number of analysed keywords	175	135
The highest number of keyword occurrences	319	249
Characteristic of the most numerous clusters	strategy/agility in general/industry	agility in general/industry

In terms of publication metrics, the most cited article in the WOS database is a 2003 article (1,614 citations) in MIS Quarterly (7.3 journal Impact Factor (2022); Q1). This also agrees with the Scopus database (the number of citations is 2,346). Of the WOS database examined, most articles are published in the International Journal of Production Economics (35) with journal metrics 19.3 CiteScore, 12 Impact Factor; International Journal of Production Research (31) with journal metrics 9.2 (2022) Impact Factor; Q1) and Benchmarking - An International Journal (23) (5.6 Impact Factor (2024); CiteScore 9.7 (2022)). In the Scopus database, most publications are in Lecture Notes In Business Information Processing (book series of Springer Nature; 101), International Journal of Production Research (23) and International Journal of Production Economics (18).

4. Conclusion

Bibliometric analysis is a useful tool in processing and detecting trends and patterns. The aim of this paper was to provide a brief insight into business agility through bibliometric analysis of articles included in WOS and Scopus as the most commonly used databases by scholars. As a result, a comparison of these databases shows that, although the databases overlap to some extent, there are some slight differences, either in terms of bibliometrics or scientometrics. Based on the keyword analysis, it is possible to note an increased interest in the issue of agility, which is linked to a number of other areas of management. To study business agility, we recommend the Scopus database.

Acknowledgement

The paper is a partial result of the grant scientific project VEGA 1/0333/22.

References

Business Agility Institute. (2023). Business Agility Report. Leading through uncertainty, 6th edition, 2023. Retrieved February 10, 2024, from https://api.businessagility.institute/storage/files/download-library/2023-11%20BAI-Business-Agility-Report-2023.pdf

Cobo, M.J., López-Herrera, A.G., Herrera-Viedma, E., & Herrera, F. (2012). SciMAT: A new Science Mapping Analysis Software Tool. *Journal of the American Society for Information Science and Technology*, 63(8), 1609-1630. https://doi.org/ 10.1002/asi.22688

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim , W. M. (2021). How to conduct a bibliometric analysis: An overview of guidelines. *Journal of Business Research*, 133, 285-296. https://doi.org/10.1016/j.jbusres.2021.04.070

Godin, B. (2006). On the origins of bibliometrics. *Scientometrics,* 68(1), 109-133. https://doi.org/10.1007/s11192-006-0086-0

Chen, Ch. (2017). Science mapping: A systematic review of the literature. *Journal of Data and Information Science*, 2(2). 1-40. https://doi.org/10.1515/jdis-2017-0006

Lyu, P., Liu, X., & Yao, T. (2023). A bibliometric analysis of literature on bibliometrics in the recent half-century. *Journal of Information Science.* https://doi.org/10.1177/01655515231191233

McBurney, M.K, & Novak, P. L. (2002). What is bibliometrics and why should you care? Proceedings. *IEEE International Professional Communication Conference, Portland, OR, USA, 108-114.* https://doi.org/10.1109/IPCC.2002.1049094

Moral-Muñoz, J. A., Herrera-Viedma, E., Santisteban-Espejo, A., & Cobo, M. J. (2020). Software tools for conducting bibliometric analysis in science: An up-to-date review. *El Profesional de la Información*, 29(1). https://doi.org/10.3145/epi.2020.ene.03

Myers, B. A., & Kahn, K. L. (2021). Practical publication metrics for academics. Clinical and Translational Science, 14(5), 1705-1712. doi:https://doi.org/10.1111/cts.13067

Van Eck, N. J., & Waltman, L. (2022). VOSviewer Manual. Manual for VOSviewer version 1.6.18.

Van Oosterhout, M., Waarts, E., & Van Hillegersberg, J. (2006). Change factors requiring agility and implications for IT. *European Journal of Information Systems*, 15, 132-145. https://doi.org/10.1057/palgrave.ejis.3000601

Yang, S., & Yuan, Q. (2017). Are Scientometrics, Informetrics, and Bibliometrics Different?. *Proceedings of the 16th International Conference on Scientometrics & Informetrics (ISSI2017); Wuhan, China, 16-20 October 2017.*

Yusuf, M., Surya, B., Menne, F., Ruslan, M., Suriani, S., & Iskandar, I. (2023). Business agility and competitive advantage of SMEs in Makassar City, Indonesia. Sustainability, 15(1). https://doi.org/10.3390/su15010627

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17829

Viability of Artificial Intelligence application for real estate valuation of Data Centers

Salvador Domínguez Gil[1], Andrea San José Cabrero[2], Antonio Sánchez Gea[3], Pilar Miguel-Sin[1], Gema Ramírez Pacheco[1]

[1]Universidad Politécnica de Madrid, Spain, [2]Universidad de Navarra, Spain, [3]SOCOTEC, Spain.

How to cite: Domínguez Gil, S.; San José Cabrero, A.; Sánchez Gea, A.; Miguel-Sin, P.; Ramírez Pacheco, G. 2024. Viability of Artificial Intelligence application for real estate valuation of Data Centers. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17829

Abstract

In the era of digitalization, Artificial Intelligence (AI) has emerged as a revolutionary tool in multiple sectors, including real estate. The ability of AI to process and analyze large volumes of data has enhancened decision-making processes on real estate activities, optimizing operations, investment strategies and assets management. However, despite its many advantages, the application of AI-based systems in the real estate sector faces significant limitations, especially in contexts of data scarcity. It is the case of emerging or minority real estate markets, such as Data Centers market.

The objective of this paper focuses on examining these limitations, highlighting the importance of considering intangible values and categorical variables, as well as the quality of the data in the evaluation of such real estate assets.

Keywords: *Real Estate assessment; Algorithmic models; Intangible value; Data quality; Data-driven decision making.*

1. Introduction

Automated Valuation Models (AVMs), commonly known as Computed Aided Mass Appraisal Systems (CAMA) have been present in real estate operations for several years now. Its application has been mainly developed towards traditional markets such as residential and offices, which occupy a large percentage of real estate assets in cities. However, these models have faced several limitations, highlitghting the applicability or confidence of the obtained results, being these related to data quality or scarcity, as well as methodology or interpretability of results, among others. In the real estate sector, most of these limitations apply to markets that are not very representartive, to markets that are underdeveloped or in the process of development, as well as to assets of unique or exclusive nature.

On the other hand, one of the main limitations of AI-based systems in the real estate sector that has arisen in recent years is the difficulty of quantifying and valuing intangible aspects of assets. Characteristics such as aesthetics, recognition, and iconography, or even potential descriptive aspects of the property foreseen in its future, are not easily amenable to direct quantification. This way, they are often evaluated in qualitative terms and through categorical variables.

Additionaly, intangible factors can significantly influence the value of a real estate assets, but their subjective and qualitative nature presents challenges for Artificial Intelligence algorithms. That is because these systems rely heavily on structured and quantifiable data, while the results are to be supported by expert judgment to incorporate inputs into decision-making.

2. Limitations of Artificial Intelligence in Real Estate

On the first hand, the scarcity of data represents a significant obstacle in the effective implementation of AI in the real estate sector. This point is common to Life Cycle Assessment (LCA) methodologies (Schneider-Marin & Lang, 2020) that are to be applied by AI. Artificial Intelligence models require large datasets to be trained and generate accurate predictions, which will depend on the type of intelligence applied and the selected model. However, in the case of unique real estate assets or in less developed markets, the availability of historical and current data may be limited. This data scarcity can lead to AI models that are not adequately informed, resulting in less reliable predictions and evaluations.

Like any property linked to economic activity, its valuation will be tied to the income it can generate throughout its life cycle, linked to Life Cycle Costing (LCC) methodologies. Therefore, the assessment is often carried through oversimplification to economic units of measurement (ISO 15686-5:2017, 2017). In this regard, the asset management process or its lease situation determines its success. Depending on its purpose and the property owner’s objectives, we will be talking about situations of more or less levels of risk that affect the financial situation of the property, as well as investors' perception of it. Additionally, established tenure model, occupancy, as well as the quality of operation are related qualifiable attributes and, therefore, new drivers with further limitations or dificulties for data structures.

There is a significant limitation with real estate valuation models as they fail to adapt to other values that are more distant from the sector. The individualized view of the property has transformed into a global analysis of physical, management, operation, and maintenance perspectives, as well as considerations around its surrounding environment, implying a shift in scenario where new variables or externalities must be included when valuing it. In this matter, the incorporation of intangible measures allows companies to increase profitability without necessarily having a commensurate increase in investment (Orhangazi, 2019), making this a fact of great interest in market research.

The quality of available data is another critical factor. Real estate information can be incomplete, outdated, and inconsistent, directly affecting the effectiveness of AI-based systems. The accuracy and reliability of AI models inherently depend on the quality of input data. Poor data management and lack of uniform standards in data collection and processing can lead to biases and errors in AI-based assessments and decisions. This shortcoming is particularly enhanced by the emergence of new methodologies of data measurement.

Furthermore, interpreting and explaining the results generated by Artificial Intelligence systems in the real estate sector poses a significant challenge, as it does in existing methodologies outside of AI. When using Life Cycle Assessment methodologies, the difficulty of understanding the Key Performance Indicators (KPIs) is highlighted often by contradictions and inconsistencies due to the lackafe of common units of measurement for these indicators (Schneider-Marin & Lang, 2020). The often "black-box" nature of AI algorithms can hinder understanding of how a particular conclusion or valuation is reached. This is particularly problematic in the real estate sector, where investment and management decisions must be based on detailed analysis and clear justifications, with risk margins and evaluations depending on the purpose of the process. The lack of transparency and explainability in AI systems can limit their utility and acceptance among industry professionals.

3. Peculiarities of a real estate asset such as Data Centers

The development of Industry 4.0, cloud computing, new internet applications and social media, will ensure continued reliance upon exponential amouns of data and processing power. This data and processing equipment needs to reside in a data centre that provides specialist environmental and security features. As a result, the data centre industry has seen high supply and demand growth in the last few years (CBRE, 2021).

The application of Artificial Intelligence in highly specialized sectors, such as data centers or high-risk biological laboratories, illustrates both the potential and limitations of these systems. In these environments, the management and evaluation of real estate assets require unique and highly specialized considerations, which may be beyond the current capabilities of commonly known AI systems and adapting expert judgment to new variables or necessary data. Site suitability assessment, for example, involves not only traditional real estate variables but also complex technical considerations related to connectivity, data security, energy efficiency, as well as industry trends and market evolution, regulations, public policies, forecasting, or risk of new requirements, etc. Similarly, high-risk biological laboratories must comply with strict safety and containment standards, requiring detailed assessments that go beyond typical real estate valuation parameters.

Among the main factors to consider in the valuation of data centers, we find the electrical power supply of the property, income and leasing or management situation, location, and the existence

of risks to the activity. The latter is a fundamental qualitative aspect when dealing with properties with very high demands from operators regarding potential floods, contamination of nearby land, or the presence of industries using hazardous materials or fertilizers (Colliers, 2023).

However, beyond the availability of land and electrical power supply at a particular geographic point, regulation from a governmental perspective, as well as political stability and sustainability, are variables to consider in the value of a data center. Since the implementation of the European Union's General Data Protection Regulation (GDPR), the construction of this type of property has increased, and consequently, any regulation requiring greater justification of value or business actions based on data will enhance its value. Another aspect to consider is sustainability, as these types of properties promote greater efficiency through new technologies. On the other hand, the carbon footprint of data centers is higher than any other real estate asset.

The International Energy Agency shows that centers use 200 TWh of electricity and generate 3.5% of the global greenhouse gas (GHG) emissions, the majority of which is utilized within the Information and Communication Technology (ITC) sector, and it is responsible for 0.3% of overall CO2 emissions (Monserrate, 2022).

4. Conclusions

Robert Sternberg, psychologist at Yale University, defines intelligence as any mental activity directed towards intentional adaptation, selection, or transformation of relevant real-world environments in one's own existence. In the case of Artificial Intelligence, it is the combination of algorithms designed with the purpose of creating machines that exhibit the same capabilities as humans (Gardner, 1987).

These algorithms, often complex and with different applicability, have been developed exponentially over the years: from linear regression models to supervised models that seek to address more common predictive or classification problems, and unsupervised algorithms that, despite not being oriented towards directly solving prediction problems, serve to interact with human rationality and other intermediate models to help solve problems in a combined manner. Additionally, recommendation algorithms or generative Artificial Intelligences have been added to this.

However, despite the multitude of existing methodologies applicable to such intelligence, data is not relevant from a quantitative aspect of mere storage but rather how it can be used. In technological processes where training and evaluation phases are fundamental, there is a quantitative need for information, but also qualitative.

Taking the latter into account, the valuation of assets such as data centers poses a challenge for the real estate sector, both in the collection of information due to the limited historical

development of these types of assets and the lack of identification of specific variables that have a direct impact on their value.

References

Bahramian, M. & Yetilmezsoy, K. (2020).Life cycle assessment of the building industry: An overview of two decades of research (1995–2018). Energy and Buildings, 219, 109917. https://doi.org/10.1016/j.enbuild.2020.109917

CBRE (2021). Spain Data Centers. How cloud interest is creating a new market.

Colliers (2023). Valoración de Data Centers.

Gardner, H. (1987). The theory of multiple intelligences. *Annals of Dyslexia*, 37, 19-35. https://doi.org/10.1007/BF02648057

ISO. (2017). ISO/DIS 15686-5:2017, Building and constructed assets — Service life planning — Part 5: Life Cycle Costing.

Monserrate, S.G. (2022). The Staggering Ecological Impacts of Computation and the Cloud. *The MIT Press Reader*.

Orhangazi, Ö. (2019). The role of intangible assets in explaining the investment–profit puzzle. In Cambridge Journal of Economics (Vol. 43, Issue 5). https://doi.org/10.1093/cje/bey046

Schneider-Marin, P. &Lang, W. (2020). Environmental costs of buildings: monetary valuation of ecological indicators for the building industry. International Journal of Life Cycle Assessment, 25(9), 1637–1659. https://doi.org/10.1007/s11367-020-01784-y

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17835

Eliciting and Retrieving the Feedback-Loop. Exploring Elicitation Interview Techniques for Detecting Algorithmic Feedback on Social Media and Cultural Consumption

Gabriella Punziano[1], Alessandro Gandini[3], Alessandro Caliandro[2], Massimo Airoldi[3], Giuseppe Michele Padricelli[1], Suania Acampa[1], Domenico Trezza[1], Noemi Crescentini[1], Ilir Rama[3]

[1]Department of Social Science, University of Naples Federico II, Italy, [2]Department of Political and Social Science, University of Pavia, Italy, [3]Department of Social and Political Science, University of Milan, Italy.

How to cite: Punziano, G.; Gandini, A.; Caliandro, A.; Airoldi, M.; Padricelli, G.M.; Acampa, S.; Trezza, D.; Crescentini, N.; Rama, I. 2024. Eliciting and Retrieving the Feedback-Loop. Exploring Elicitation Interview Techniques for Detecting Algorithmic Feedback on Social Media and Cultural Consumption. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17835

Abstract

This article introduces elicitative interviewing techniques in the context of algorithmic feedback detection on social media about cultural consumption. This article presents elicitation interviewing methods to identify algorithmic feedback concerning cultural consumption on social media. The initial section will clarify the notion of influence in algorithm-driven consumption decisions on these platforms. The second part will underscore the necessity for finely nuanced qualitative methodologies to dissect the conceptual facets essential for analysis within such contexts of influence and dynamics. The main interviewing techniques for finalizing data collection with this intent will then be reviewed. The third part will present an example of a survey instrument that uses the elicitation component to achieve the essence of the feedback-loop between algorithms and cultural consumption choices that underlie the PRIN ALGOFEED survey. Finally, this detection phase's placement within the project and its role as an enhancer of the preceding collection and analysis stages will be elucidated, emphasizing the benefits of this decision and the potential pitfalls that necessitate proper attention and scrutiny.

Keywords: *Algofeed; Algorithmic Recommendations; Feadback-loop; Qualitative digital research; Elicitative interview*

1. Introduction: Algorithmic Feedback on Social Media and Cultural Consumption

In the context of contemporary sociological studies, it is observed that the digital scenario has exerted an incisive transformation on how individuals structure their daily existence, articulate the expression of personal identity, and participate in the production, dissemination, and reception of belief systems, bodies of knowledge and preferential orientations. In this context, the influence exerted by social media, distinguished by its ability to facilitate interactions, promote integration and achieve synchronization with various media formats, emerges with preponderance (Weingartner, 2021). This capacity has progressively given social platforms a pivotal role in aggregating and researching interests manifested through cultural consumption practices. Searching, viewing, and sharing activities on social media catalyse processes and dynamics that focus discovery, critical analysis, and debate on cinematic, literary, musical, and other manifestations of the cultural landscape, thereby reconfiguring how culture is consumed, interpreted, and negotiated in the digital age as well as the generative modes of tastes, fashions, and trends.

Social researchers have been exploring various realms of inquiry extensively. Consider, for instance, the contextualization of digital cognitive spaces where interactions take place (Boccia Artieri, 2012; Bennato, 2021), the mechanisms of content mediation and remediation (Couldry, 2013), or contemporary art forms that are either transposed or directly originated online (De Seriis, 2008). The current focus of social research is the complex structures of digital ecosystems that are no longer configured solely through user-generated content (UGC), but are also characterized by sociotechnical artifacts that operate according to algorithmic logic. The latter plays a crucial role in digital platforms' filtering and stratification procedures in content distribution. This dynamic highlights a significant evolution in digital media's information and communication architecture, marking a shift from a predominantly participatory model to one structurally mediated by algorithmic systems. These systems influence not only the visibility and accessibility of information but also the construction of hierarchies of relevance and meaning within the digital space, thus affecting the formation of public opinion and cultural consumption practices.

Contemporary sociological reflection in recent years aims to analyze the repercussions of algorithmic mechanisms on society, focusing on their ability to influence how social interaction, the configuration of cultural identities, and the modulation of power dynamics within the current media ecosystem are affected. This highlights how algorithmic architectures are not neutral but act as active mediators that can foster, limit or direct the visibility of specific content, narratives or voices, reflecting and potentially reinforcing pre-existing power structures and social inequalities.

In the context of technologically mediated sociocultural dynamics, it is observed that algorithms act as catalysts in broadening consumers' horizons, subjecting them to a curated set of content consonant with their prior tastes (Airoldi & Rokka, 2022). In parallel, making use of the digital traces left by consumers, machine learning mechanisms iteratively elaborate their output strategies, pursuing the goal of synchronizing future proposals with the perspectives and expectations of recipients (Brinckmann, 2023). This continuous feedback and adjustment process between algorithms and human behavior highlights a co-evolution of technological systems and consumption practices, in which individual and collective preferences are simultaneously reflected and shaped through interaction with advanced information systems.

Algorithmic feedback can significantly impact online users' behavior and perceptions, influencing what content is viewed, what other users interact with, and what information is assimilated. However, debate persists regarding the ethical and social implications of algorithmic feedback, including privacy, manipulation of public opinion, the creation of information echo chambers, and guiding consumer choices. The fundamental explanation is that the algorithmic feedback chain is powered by factors such as those who instigate specific behaviors, including consumption behaviors, by targeting sponsored elements and by users' recursive decisions. Once integrated into the system, discerning who influences what and to what extent within the generated vortex becomes challenging. Does this align with the feedback-loop concept introduced by Airoldi (2021), and who is the focus of our discussion

Digital pervasiveness in the daily lives of every segment of society over the past 20 years has been a turning point. Therefore social research, too, has yet to be caught unprepared, facing multiple reflections on methodological opportunities helpful in understanding how to adapt research actions and techniques (Punziano & Delli Paoli, 2021). It is crucial to move beyond applying a single method based on quantitative approaches and a standardized view of the phenomenon to better understand social platforms' impact on cultural consumption. To obtain a comprehensive view of cultural consumption processes, it is relevant to go beyond the need for retrievable information that can refer to all types of users (such as time spent online, amount of connections with other users, main content searched, etc.). It may be helpful to investigate the emotions, perceptions, and experiences that users are addressing. In light of this, qualitative methods may be the most suitable option, in a mixed methods research framework. Building upon these premises, this article aims to delve into the methodological construction of the elicitation interview technique, namely the qualitative detection of algorithmic feedback on social media platforms, specifically TikTok and YouTube, within the framework of the PRIN 2022 project “Algofeed”. It highlights the construction of the elicitation instrument as the third phase of the mixed methodological approach and explores its potential implications on the outcomes of the Algofeed research. The subsequent sections will elucidate the methodological approach, focusing on elucidative interviewing techniques, and exemplify their application in uncovering the dynamics of algorithmic feedback loops.

2. The qualitative approach on complex digital phenomena: eliciting, solicit, request

Qualitative digital research refers to the use of qualitative research methods within the context of digital environments. It involves studying human behaviour, attitudes, and experiences online using various digital tools and platforms. Qualitative research focuses on in-depth understanding of individuals' perceptions, motivations, and behaviours. This involves interviews, focus groups, participant observation, and content analysis. In the digital environment, qualitative digital research takes place within scenarios such as social media platforms, online communities, forums, websites, mobile apps, and digital communication channels where essentially researchers collect data from digital sources, which may include user-generated content, social media posts, comments, messages, multimedia content (such as images and videos), and interactions within online communities. Researchers use qualitative analysis techniques in analyzing these materials to make sense of the digital data collected. This may involve coding, thematic analysis, discourse analysis, and other qualitative data analysis methods to identify patterns, themes, and insights. Of fundamental importance is the contextual understanding, for which qualitative digital research emphasizes understanding the context in which digital interactions occur, fully assuming the non-neutrality of the digital scenario (Rogers, 2009). This includes considering factors such as the online platform's features, the social dynamics of online communities, and the cultural norms shaping digital behaviours. In this research path, researchers must address ethical considerations specific to digital research, such as privacy concerns, informed consent, data security, and the implications of studying online communities and individuals in digital spaces. An interdisciplinary approach is often involved, drawing on fields such as sociology, psychology, anthropology, communication studies, information science, and digital humanities to provide a comprehensive understanding of digital phenomena but more importantly to have at their disposal mining skills, data collection, and interpretive sensitivity that are unlikely to be the assets of individual scholars. So much so that the application areas for qualitative digital research can be passed through various domains, including but not limited to consumer behavior analysis, social media studies, online community research, digital ethnography, and user experience (UX) research. Overall, qualitative digital research offers valuable insights into the complex interactions and behaviours within digital environments, contributing to our understanding of the digital world and its impact on society (Bryda & Costa, 2023).

But how can the field of algorithmic feedback investigation benefit in the study of the influence of social media on in cultural consumption? Among recent analytical frontiers, there has been a flood of literature regarding a specific approach to get to the heart of the ontology of the consumption choices of users embedded in social networks, and that is the field of elicitation interviewing. An elicitative interview is a qualitative research method used to gather information from participants by prompting them to share their experiences, perspectives, and insights in a

structured yet open-ended manner. The term elicitative refers to drawing out or eliciting responses from participants using open-ended questions and prompts that encourage participants to provide detailed and spontaneous responses. Unlike structured interviews with fixed questions, elicitative interviews allow for flexibility and exploration of participants' thoughts and experiences. Elicitation aims to enrich conventional methodological approaches and has its foundations in anthropology and visual sociology. Within this framework, elicit interviews are not merely limited to the collection of data considered in a sense as objective, but rather aspire to grant the researcher the opportunity to penetrate layers of meaning and interpretation that transcend what is manifest on the surface. In parallel, such interviews offer respondents the privilege of articulating their thoughts, opinions and perceptions comprehensively and contextually relevantly. This approach fosters a co-construction of data through interaction, as highlighted by Salvini (2015), assuming that the researcher has a keen understanding of relational dynamics and communicative strategies for effective implementation. Essentially, the elicitation technique finds its foundation in the production of discourse from artifacts, usually through dialogic interaction between participants and the researcher, centered around one or more elements that often take the form of visual representations, in the context of empirical information generation and, more traditionally, in the context of qualitative interviewing (Giorgi et al., 2021). The primary goal of elicitative interviews is to explore participants' perspectives, beliefs, attitudes, and experiences related to a specific topic or research question. Researchers seek to understand the richness and complexity of participants' viewpoints through probing and follow-up questions. This kind of interview prioritize the participant's viewpoint and experience, so are defined participant-centered. Researchers aim to create a comfortable, non-threatening environment encouraging participants to share openly and honestly. This may involve building rapport, active listening, and demonstrating empathy and respect for participants' perspectives. On the side of the data collection, elicitative interviews are a form of qualitative data collection that generate rich, in-depth data that can provide insights into individuals' motivations, behaviours, and perceptions identified by the researchers in the shape of patterns, themes, and key findings. This specific instrument allows for flexibility and adaptability in the interview process. Researchers may adjust their approach based on the participant's responses, probing further into areas of interest or exploration. This flexibility enables researchers to uncover unexpected insights and nuances. Among the most common elicitation techniques are photography (Harper, 2002) and photovoice (Wang, 1999) delineating these tools as effective strategies in promoting a process of incorporation. In this sense, through the mediation of visual sensory stimuli, it is possible to facilitate access to symbolic and material constructs related to personal identity and lived experiences, allowing for decoding them in symbolic and concrete keys (Vacchelli, 2018). On and through the digital, conducting these kinds of interviews and administering these kinds of stimuli is most benefited by the mediation of the screen on which images and audiovisual materials can pass easily. Another type of elicitative interview technique used is the semi-structured interview enriched with perceptual

stimuli. This approach allows interviewers to have a fluid interaction that is not strictly composed of a series of questions and topics to be addressed to which is combined with the use of constructed stimuli such as questions or visual and audiovisual graphic stimuli aimed at transporting the interview into the vital way it is intended to be explored. However, it leaves room for flexibility and spontaneity during the interaction. This allows interviewees to express themselves freely, without feeling constrained by a rigid schema, and for researchers to delve into themes that emerge during the conversation. The respondent is prompted to question his or her choices critically, allowing a deep understanding of the dynamics underlying decision-making in complex algorithmically mediated contexts. The elicitation interview is also not infrequently associated with the phenomenological interview. This concerns exploring participants' lived experiences, channeling attention to the perceptions, emotions and meanings individuals attribute to their social media interactions in relation to cultural consumption. This approach is based on the theoretical assumption that understanding social phenomena requires an immersion into the inner perspectives of subjects in order to unravel the essential structures of lived experience (Husserl, 1931). Following Moustakas (1994), the primary objective of the phenomenological interview is thus to access the "pure consciousness" of the participants, allowing researchers to grasp the richness and complexity of individual experience in an unmediated manner. Through this lens, the phenomenon of cultural consumption on social media is investigated not only in its outward manifestation but, more importantly, in its intrinsic and experiential dimensions, emphasizing the importance of contextualising individual experiences within their specific plots of meaning. The use of the phenomenological interview technique in the context of social sciences allows for the exploration and elucidation of the complex dynamics underlying cultural consumption behaviors, thus providing meaningful insights for deciphering practices in digital environments (Van Manen, 1990). Adherence to ethical principles is a cornerstone in sociological research, especially in contexts that involve direct interaction with participants as is precisely the case with elicitative and phenomenological interviews. Informed consent, privacy protection, and personal data protection emerge not only as legal obligations but as fundamental moral imperatives that uphold the integrity of the research process (Bryman, 2016). Furthermore, ethical management of research requires an empathetic and sensitive approach on the part of the researcher, who must strive to build a safe and welcoming environment that facilitates the free expression of participants, while ensuring that their experiences and perceptions are treated with the utmost respect and consideration (Ellis, 2007). To enrich the final goals of the research, in this field researchers typically use purposeful sampling techniques to select participants who can provide relevant and diverse perspectives on the research topic. This ensures that the data collected during elicitative interviews are representative and meaningful. In other cases, this type of interviewing is used in mixed methods projects where interviews are conducted as qualitative follow-ups on respondent/user/subject profiles deducted from previous quantitative research phases (as well as a survey or an experiment). The individuals selected will correspond exactly to the identified

profiles and will be representative of them, although they will be chosen incidentally from all the individuals who fall into that specific profile. As will be seen in the next paragraph, this is precisely the mode of use described in our research example. The elicit interviews require the researchers to reflect on their own biases, assumptions, and positionality throughout the research process. Practicing reflexivity helps researchers acknowledge their influence on the interview process and interpret participants' responses in context (Giorgi et al., 2021).

3. Qualitative detection of the loop: an example of elicitative interview

In the context of the PRIN 2022 project financed by *European Union – Next Generation EU*" – "Feedback culture: assessing the effects of algorithmic recommendations on platformized consumption – ALGOFEED", whom the main objective is to explore the sociocultural effects of feedback loops based on digital platforms (particularly TikTok and YouTube) it was utilized a prototype of elicitation interview. The methodological approach, mixed at the basis, involves two preparatory phases before conducting interviews: a preliminary survey on user perception of the algorithmic component in the cultural consumption choices and a self-tracking to detect the effective recommendation paths and user profiles. These stages are preparatory to using of semi-structured interviews, which are useful for achieving various purposes, such as a greater focus on the proposed subject (Bichi, 2007).

The key dimensions structured for the creation of the interview guide, as illustrated in Figure 1, include 1) platforms and cultural consumption; 2) content filtering; 3) automated decision-making processes; 4) algorithmic persuasion, referring to the awareness of the potential of algorithms in shaping users' behavioural choices and content consumption patterns. Finally, 5) human-algorithmic interaction, focusing on the analysis of the dynamic interaction between individuals and algorithms in digital contexts. These conceptual dimensions are fundamental for delineating the research scope and guiding the formulation of survey questions in the context of the ALGOFEED project.

To explore the dimension related to understanding and individual reactions towards content filtering, it is helpful to consider the interviewees' awareness of using algorithms in content personalization and their reflections on the dynamics of this personalization. , In this phase, elicitation is structured through a dynamic interaction between the researcher and the interviewee and is based on the generation of words that arise from visual sensory stimuli and access representations of one's identity or experience. Specifically, in this work, both engage in viewing the first 10 videos that appear in the respondent's TikTok feed and the first 6 video suggestions that appear on YouTube.

In this way, it is possible to discuss the respondents' algorithmic awareness, the customisation of the recommended content according to the interactions that take place online. In fact,

following an outline of a detailed account of their daily routine, highlighting the specific moments when interaction with the YouTube and TikTok platforms occurs.

In an attempt to further explore the relationship between digital platforms and users' cultural consumption, we make use of the elicitative interview technique. The latter aims to explore participants' daily habits and their interaction with platforms such as YouTube and TikTok and allow us to investigate our objective, which is related to understanding how these platforms are integrated into the everyday cultural context, influencing it and shaping its dynamics.

This will allow the identification of consumption patterns, usage frequency, specific contexts of use, and how recommendation algorithms intertwine with individuals' daily lives, shaping their cultural choices and preferences. These will be just some of the elicitation techniques aimed at facilitating dynamic exchange between the researcher and the interviewee, thereby promoting the acquisition of in-depth knowledge.

4. The dimensions of the elicitation interview. Methodological notes and insights

The tool's framework addresses five dimensions of the digital sphere concerning algorithmic awareness, adopting the perspective of the elicitation technique, which involves the use of stimuli that allow the respondent to anchor themselves to the object of research. In our case, the use of digital content such as TikTok and YouTube videos - and hence their related recommendation outputs - represents an innovative approach enabling the simulation of contexts where the algorithmic element could intervene significantly. Particularly, as shown by the visual model proposed (Figure 1), the interview outline encompasses specific information. The first dimension concerns platforms and cultural consumption, while the second focuses on content filtering. The former pertains the type and frequency of digital platforms usage; the latter dimension is particularly associated with awareness regarding algorithm usage to personalize content recommendations based on online data. Both dimensions are explored through the elicitation of "scrolling", during which the researcher prompts the respondent to examine together the first 10 videos viewed on TikTok and the first 6 video recommendations on YouTube. This approach encourages reflection on the visual material observed and initiates discussion on the frequency of certain types of videos.

In this context, the interview's elicitation aims to incorporate elements of innovation, creativity and depth. This is enabled using digital audiovisual stimuli to immerse the interviewee in realistic situations and thus capture their perceptions in real time. Secondly, stimulating contexts where algorithmic recommendations are evident helps the interviewer to understand the perceived influence of the algorithm on cultural consumption.

This initial exploratory research phase of elicitation thus helps the researcher obtain valuable initial frameworks useful for discerning how the feedback loop effect is shaped. This involves

detecting the types of content reaching users based on their individual browsing patterns and online presence. Consequently, it is important to delve deeper to understand whether, and if so how, users are aware of the dynamics and processes that lead to such recommendations. The involvement of interviewees in envisioning practical and realistic scenarios, such as making decisions based on online content or assessing the reliability of reviews when shopping, leads to the third dimension of the interview related to the automated decision-making processes. It investigates whether the respondent has ever made decisions based on online content, engaging them in potentially relevant scenarios and practices adopted online, such as planning trips or discovering new recipes due to consumed digital platform content, or the importance of leaving online reviews when shopping and other online research practices.

The elicitation support, which prompts personal experiences of the interviewees, is similarly observed in the fourth dimension. Relying on personal experiences allows the exploration of the respondent's emotional and cognitive reactions in similar situations, shedding light about the algorithmic persuasion, i.e. whether there is awareness of algorithms potentially influencing users' behavior and choices. This, evidence emerging from this interview segment partially clarifies the dynamics of the feedback loop and how it is processed and developed.

The fourth dimension pertains to algorithmic persuasion, addressing whether there is awareness that algorithms may or may not influence users' behaviour and choices. Both dimensions are supported by the elicitation of reference to personal experiences by the researcher to stimulate discussion. Lastly, the fifth dimension of the interview aims to illuminate users' awareness of the feedback loop effect. It focuses on human-algorithm interaction. It involves asking if the interviewee has ever read the terms of use of online platforms and whether they found it useful. Discussion also delves into the feelings evoked by online content, such as emotions and personal reactions.

5. Conclusion

The paper aimed to outline the key technical phases involved in constructing the elicitation interview. This technique represents the final step of a broader national research project and serves as a cohesive element with two other quantitative stages within a fully mixed-methods framework. This research phase has introduced several innovative elements. Although the technique used belongs to the broader family of interviews - widely used techniques in the humanities and social sciences - the elicitation tool, as observed, has generated methodological innovation that aligns well not only with the short-term objectives of this third phase - namely, a deep understanding of the relationship between the user and algorithmic recommendation - but also with the long-term objectives of the entire research project. In this case, the elicitation interview has been a fundamental tool for delving into the experiences involving cultural

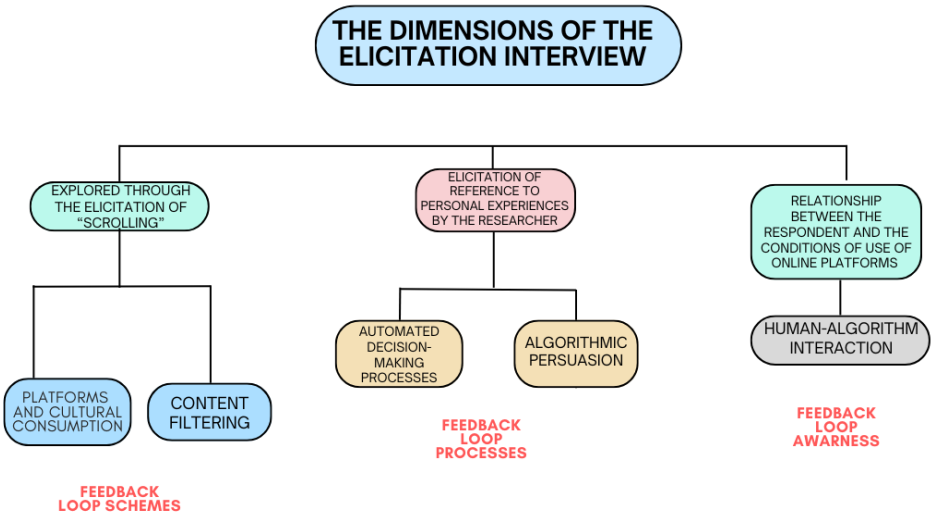


Figure 1. Visual model of interview dimensions and the chosen elicitation

consumption and how it fuels, and is fueled by, the automated processes of the Network. Indeed, "scrolling", the researcher's personal experiences, and the deepening of the conditions of use of the platforms under investigation have represented three very effective strategies for bringing to light the interviewee's digital experiences and investigating their action patterns in heavily "algorithmized" contexts, and whether these are somehow oriented by default recommended patterns. However, several challenges loom in the realm of data collection and analysis, including managing process disparities, drawing accurate insights, conducting qualitative aspects reliant on researchers' skills, and navigating various operational, analytical, and interpretative languages, necessitating robust team cohesion and clarity of objectives. Addressing these challenges falls upon the application phase, where they will be further problematized and examined.

References

Airoldi, M. (2021). *Machine habitus: Toward a sociology of algorithms*. John Wiley & Sons.

Airoldi, M., & Rokka, J. (2022). Algorithmic consumer culture. *Consumption Markets & Culture*, *25*(5), 411-428.

Bennato, D. (2021). The digital traces' diamond. a proposal to put together a quantitative approach, interpretive methods, and computational tools. *Italian Sociological Review*, *11*(4S), 207-207.

Bichi, R. (2005). *La conduzione delle interviste nella ricerca sociale* (pp. 1-216). Roma: Carocci Editore.

Boccia Artieri, G. (2012). *Stati di connessione. Pubblici, cittadini e consumatori nella (Social) Network Society* (Vol. 1097). FrancoAngeli.

Brinkmann, L., Baumann, F., Bonnefon, J. F., Derex, M., Müller, T. F., Nussberger, A. M., ... & Rahwan, I. (2023). Machine culture. *Nature Human Behaviour*, *7*(11), 1855-1868.

Bryda, G., & Costa, A. P. (2023). Qualitative research in digital era: innovations, methodologies and collaborations. *Social Sciences*, *12*(10), 570.

Bryman, A. (2016). *Social research methods* (5th ed.). Oxford, UK: Oxford University Press.

Couldry, N. (2013). *Why media ethics still matters. Global media ethics: Problems and perspectives*, 13-28.

Deseriis, M., & Marano, G. (2008*). Net Art. L'arte della connessione*. Milano, Shake Edizioni.

Ellis, C. (2007). *Telling secrets, revealing lives: Relational ethics in research with intimate others*. Qualitative Inquiry, 13(1), 3-29.

Giorgi, A., Pizzolati, M., Vacchelli, E. (2021). *Metodi creativi per la ricerca sociale. Contesto, pratiche, strumenti*, Bologna, Il Mulino.

Harper, D. (2002). Talking about pictures: a case of photo elicitation, in Visual Studies, vol. 17, n.1, pp. 13-26

Husserl, E. (1931). *Meditazioni cartesiane*. Parigi: Félix Alcan.

Israel, M., & Hay, I. (2006). *Research ethics for social scientists*. London, UK: Sage Publications.

Moustakas, C. (1994). *Phenomenological research methods*. Thousand Oaks, CA: Sage Publications.

Punziano, G., & Delli Paoli, A. (Eds.). (2021). *Handbook of research on advanced research methodologies for a digital society*. IGI Global.

Rogers, R. (2009). *The end of the virtual: Digital methods* (Vol. 339). Amsterdam University Press.

Salvini, A. (2015). *Percorsi di analisi dei dati qualitativi.* Torino, Utet.

Vacchelli, E. (2013). *Embodied research in migration studies: using creative and participatory approaches*, Bristol, Policy.

Van Manen, M. (1990). *Researching lived experience: Human science for an action sensitive pedagogy*. Albany, NY: State University of New York Press.

Wang, C. C. (1999). Photovoice: a partecipatory action research strategy applied to women's health, in Journal of Women's Health, vol. 8, n.2, pp.185-192.

Weingartner, S. (2021). Digital omnivores? How digital media reinforce social inequalities in cultural consumption. *New Media & Society*, *23*(11), 3370-3390.

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17834

The Algofeed project. A methodological proposal to assessing the effects of algorithmic recommendations on platformized consumption

Gabriella Punziano[1], Alessandro Gandini[3], Alessandro Caliandro[2], Massimo Airoldi[3], Giuseppe Michele Padricelli[1], Suania Acampa[1], Domenico Trezza[1], Noemi Crescentini[1], Ilir Rama[3]

[1] Department of Social Science, University of Naples Federico II, Italy, [2]Department of Political and Social Science, University of Pavia, Italy, [3] Department of Social and Political Science, University of Milan, Italy.

How to cite: Punziano, G.; Gandini, A.; Caliandro, A.; Airoldi, M.; Padricelli, G.M.; Acampa, S.; Trezza, D.; Crescentini, N.; Rama, I. 2024. The Algofeed project. A methodological proposal to assessing the effects of algorithmic recommendations on platformized consumption.. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17834

Abstract

This paper outlines the construction phase of the consumption profiling, algorithmic awareness, and digital literacy tool within the ALGOFEED project. ALGOFEED, funded as an Italian Research Project of Significant National Interest (PRIN), seeks to illuminate the socio-cultural impacts of platform-based feedback loops while developing a novel methodological and theoretical framework for the sociological examination of consumer culture algorithms. The project aims to produce unique empirical insights into cultural consumption patterns on digital platforms in Italy. Specifically, ALGOFEED investigates the effects of automated content recommendations on platforms like TikTok and YouTube on individual and collective cultural consumption behaviours over time. The primary objectives include acquiring distinctive empirical findings on platform-driven cultural consumption in Italy, advancing a fresh methodological and theoretical approach for studying algorithmic consumer culture sociologically, and fostering algorithmic awareness among Italian consumers through impactful diffusion activities grounded in a transformative paradigm.

In this study, the authors introduce the empirical tool designed to detect the feedback loop and define its components: digital skills, Platform Usage Type, and Algorithmic Awareness.

Keywords: *Algofeed; Algorithmic Recommendations; Algorithmic Awareness; Digital Skills; Cultural Consumption*

1. Introduction

The ALGOFEED project is an Italian Research Projects of Significant National Interest (PRIN) which has the general objective of shedding light on the socio-cultural effects of platform-based feedback loops, developing a new methodological and theoretical framework for the sociological study of consumer culture algorithms capable of producing unique empirical results on the cultural consumption platform in Italy. Feedback can be defined as "the property of being able to adjust future conduct by past performance" (Wiener 1989: 33). In the case of interactions between platform users and recommender algorithms, these feedback-based learning happens on algorithmic outputs expose consumers to selections of content they "may also like"; on the other hand, based on consumers' input data, machine learning systems iteratively decide how to update their future outputs, aiming to get aligned with consumers' expectations. This dynamic process of mutual influence between recommender algorithms and platform consumers will likely establish "feedback loops" (Jiang et al., 2019; figure 1). The social science literature has widely discussed platform-based feedback loops. Research has mostly focused on how online algorithms risk producing a polarized public opinion (Moller-Hartley et al., 2021), often overemphasizing the power of algorithms in a technologically deterministic fashion (Bruns, 2019). Conversely, the active role of the "human in the loop" has received less attention – apart from recent works highlighting how platform users may "resist" algorithmic power and showcase various levels of "algorithmic awareness" correlated with socio-demographic variables (Gran et al. 2020). Recent articles stress how the effects of platform-based feedback loops on online users go well beyond the political sphere, and directly concern consumer culture and habits more broadly intended (Fourcade and Johns, 2020). Scholars in the context of cultural consumption studies have noted how recursive interactions between consumers and recommender systems are likely to strengthen past consumption patterns, eventually "normalizing" them (Hallinan and Striphas 2016). This techno-social process is believed to induce the unaware adaptation of consumer tastes and identities to automated recommendations. Yet, sociological research still rests, for the most part, on theoretical speculations. This lack of empirical evidence on platforms' "feedback culture" is due to epistemological and methodological limitations: platform algorithms are opaque "black boxes" that are "immune from scrutiny" (Pasquale, 2015) due to both technical and corporate reasons. Moreover, algorithmic outputs are highly personalized and changeable over time, and thus very difficult to track. In the end, users have a limited awareness of algorithmic systems' activities, further complicating the study of their interactions with them. For all these reasons, scholars and practitioners have frequently launched calls for "opening the black box" and making platform algorithms accountable (Ananny and Crawford 2018). Considering this background, the main aim of this research project is to fill this empirical gap. How does the automated content recommendation on digital platforms affect users' individual and collective consumption patterns over time? To answer this research question, a mixed methods strategy will shed light

on the socio-cultural effects of platform-based feedback loops. Given the breadth of the topic addressed, preparing a plan for empirical advancement involving a case study limited to the Italian context was necessary. In this regard, attention focused specifically on platformed entertainment and music consumption feedback loops. From a methodological standpoint, the research path articulates through a sequential mixed-methods research design (Teddlie & Tashakkori, 2011) aimed at empirically investigating a) the joint evolution of individual entertainment and music consumption trajectories and personalized algorithmic recommendations over time, b) aggregate content recommendation trends at the platform level, and c) Italian consumers' awareness and understandings of platform-based recommendation systems. To better understand how algorithmic systems recursively shape cultural consumption, longitudinal and cross-platform examinations (Rogers 2017) capable of grasping at once individual consumers' trajectories and understandings and the macro unfolding of consumption and recommendation trends (Airoldi 2021) are required. The research focuses on two widespread video sharing platforms presenting personalized content recommendations, that is, YouTube and TikTok. The mixed-methods research design combines three concatenated operational steps of data collection: 1) a short pre-tracking survey, 2) longitudinal digital tracking, and 3) qualitative follow-up interviews. Illustrated the general methodological path of the project, in this work, we will specifically present point 1, which is the empirical tool for detecting the feedback loop and the dimension of the tool. The three dimensions are: Digital Skills, Platform Usage Type and Algorithmic Awareness.

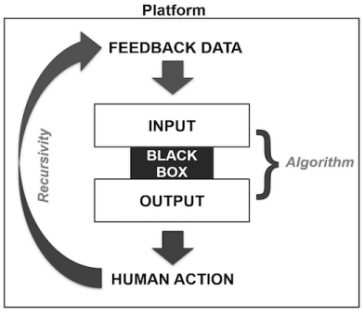


Figure 1 - Feedback loops. Source: our elaboration

2. The construction of the pre-tracking survey questionnaire

The construction of the questionnaire is aimed at profiling the sample across three specific dimensions of significance: the first concerns Digital Skills (DS), a framework on digital competencies widely experimented in literature: despite extensive reference to some existing scales in the literature, many indicators have been custom-built or modified for the research project. The second operationalizes dimensions related to the use of social platforms and their cultural consumption (Platform Usage Type, PUT). The third macro-dimension pertains to

Algorithmic Awareness (AA), specifically the questionnaire section dedicated to questions probing individuals' understanding of the algorithmic processes governing the Internet and social media platforms. This latter area, originally encompassed within the DS domain, has been separated from it conceptually because it serves as a "bridge" between skills and usage experiences. Therefore, as described later, a dedicated section has been allocated to it in the document. The empirical advancement concerning the first dimension of analysis is aimed at describing the framework on digital competencies with a specific focus on the operational phases in defining the universe of skills that users possess to interact with the world of social media; the second aims to describe the operational steps that led to the operational definition of sub-dimensions and indicators regarding the type of platform usage and media consumption practices; finally, the third concerns the description of the dimensions put in place to identify the relationship between the user and algorithmic processes, across different aspects: cognitive, procedural, behavioural, and affective. The objective is to construct the type of user defined by the experience with social networks, particularly the use of the platforms under study, TikTok and YouTube.

2.1. The construction of the pre-tracking survey questionnaire

The Strategic Program for the Digital Decade guides Europe's digital transformation establishes objectives in digital skills, digital infrastructure, digitalization of businesses, and public services. With particular reference to digital skills, the European framework DIGCOMP (figure 2) establishes a standard reference for digital skills by identifying and describing what it means to be competent in using digital technologies. DIGCOMP outlines key competencies essential for effective interaction with technology in various areas of life, such as education, work, and social participation. It aims to promote and enhance digital skills at all levels. According to the European framework, digital competence is "the ability to use digital technologies confidently and critically". According to this framework, and following Vuorikari, Kluzer and Punie (2022) digital skills are divided into four main areas (see figure 2).

Figure 2 – Digital Competence (soure DigiCompEdu 2.21)

A summary is shown in the table below.

Table 1. Areas and indicators of the dimension DS

Content Creation Skills	**Digital Communication and Interaction Processes**	**Security**
1. Creation and Modification Of Content Produced By Third Parties: • Ability to create and modify content such as texts, images, videos, music. • Level of k nowledge of software tools for content editing (e.g., text editor, video and image editing software). • Frequency of modifying third-party content (e.g., for work, hobbies). • Level Of knowledge of licenses applicable to the use of digital content (e.g., creative commons licenses).	*1. Creation and management of a profile/account on digital platforms:* • Frequency of creating or updating profiles on various digital platforms (e.g., social media, forums). • Level of detail and care in profile maintenance (e.g., type of personal information entered in the profile, frequency of publication, type of content posted, periodic profile updates). • Knowledge/Use of profile privacy and security settings (e.g., profile visibility controls, post privacy settings, contact and friend management, block and blacklist: data sharing controls with third-party apps; location settings, cookie and tracking management, etc.).	*1. Knowledge of online security practices and data protection:* • Level of knowledge of good online security practices (secure passwords, encryption, safe browsing, phishing recognition, etc.). • Browsing modes (private or incognito offered by browsers). • Frequency of applying online security practices in daily life (e.g., use of VPNs, password managers, website certificate verification). • Ability to implement personal data protection measures (e.g., encryption, backup, privacy settings). • Identification of potential risks (such as phishing messages).
	2. Knowledge of good practices in online communication: • Knowledge of netiquette rules (evaluable through concrete examples such as: do you recognize yourself in this sentence?).	

[1] Available at: https://publications.jrc.ec.europa.eu/repository/handle/JRC128415

The areas of this first dimension are addressed in the survey form through specific questions operativized with accurate scaling techniques. Using Likert scales, participants are first asked how frequently they browse the internet, which social platforms they use, and for what purposes. Moreover, through several questions, participants are asked to self-assess their proficiency in typical operations performed while browsing the internet. Also, within the dimension related to digital skills, participants are asked, through simulations of content consultation, to distinguish between messages created by individuals or bots, and to judge which types of behaviors they consider appropriate to adopt on social media.

2.1.1 Platform Usage Type

The construction of questions regarding the platform's type of use and cultural consumption is inspired by the work of Boyd and Ellison (2007) on the structure, objectives, and user types of Social Networks. The articulation of dimensions and their respective indicators in this specific section took into consideration two critical concepts in the study of social media and related practices: "contexts" and "users" (Bennato 2008; Boccia Artieri and Marinelli, 2018). The "contexts" represent the usage environments of the users and are linked to the tools of users' digital experience and the time spent. Essentially, it concerns their presence on Social Networks (General Platform Presence) and, specifically in the research context, their usage experience of YT and TT platforms (Specific Platform Presence). In the former, indicators are defined primarily to detect the essential characteristics of the user: in addition to the device usually used to access online platforms, they are asked about the main social networks they are subscribed to, the type of subscription activated, and the possible extent of expenditure incurred for content consumption. The second dimension is defined by indicators (Table 2) that address this aspect more specifically, focusing precisely on the two platforms of interest. The second macro-dimension of the questionnaire ("users") is oriented toward users and their modes of consuming and creating content and managing social networks. Specifically, three dimensions have been defined: Interactions, Networking, and Contents. The first investigates the prevalent modes of interaction with content produced by others, or towards preferred and less appreciated content. This would also allow observing some proxies related to the phenomenon of filter bubbles, presumably characterized by attitudes of constant acquiescence (Spohr, 2017). The second dimension deals with the breadth of contact networks on the two platforms, inbound (followers) and outbound (pages, users, and channels "followed"). The third dimension pertains to Contents, i.e., the mere enjoyment about prevalent themes and the creation - if any - of content, focusing on the propensity to realize or consume viral content. In this regard, a question about the user's role on the platform through a continuum from "simple content observer" to "professional content creator" better clarifies this semantic dimension.

Table 2. Areas and indicators of the dimension PUT

Contexts of Use	Users (The "Prosumer" User and Their Network)
1 General Platform Presence • Most used devices for browsing • Purpose of Internet usage • Most used Social Networks (SN) *2 Specific Platform Presence* • Type of YT/TT Account • Membership seniority • YT/TT access time • Time spent	*1 Interactions* • Primary mode of interaction with content (like, share, comment, private resharing) • Sentiment of interaction (support, neutral observation, criticism, conflict, etc.) *2 Networking Network breadth* • Number of Followers • Number of Following • Number of Channels followed *3 PCC Contents* • Prevalent consumption categories • Prevalent creation categories • Virality: Engagement for viral content (consumption/creation)

The operational definition of the described indicators leads to the development of questions within the questionnaire that focus on the longevity of users' social media presence on the platforms chosen as case studies (YouTube and TikTok). Through scaling techniques, participants are asked if they have an active account on both platforms, how frequently they access them, and how often they produce content. They are also asked about the number of followers per platform and the type of content they view or produce (video streaming, podcasts, vlogs, challenges, etc.) by theme (politics and society, TV series and shows, travel, wellness, recipes, cars, etc.).

2.1.2 Awareness Algorithmic

Algorithmic awareness (AA) is a concept that refers to individuals' understanding and awareness of how algorithms operate and are employed in various digital and social contexts.

In a single set of questions operationally defined through a Likert scale, questionnaire participants are asked to indicate their level of agreement with the characterizations of the indicator presented in Table 3; the works of Zarouali, Boerman, de Vreese (2021) and Felaco (2022) were used on AA dimensions which can be articulated as follows:

Table 3. Areas and indicators of the dimension AA

cognitive dimension	procedural dimension	behavioral dimension	affective dimension
1. Awareness of content filtering operation. •Algorithms are used to recommend multimedia content on the platform. •Algorithms are used to prioritize certain multimedia content over others. •Algorithms are used to personalize specific content on the platform.	*1. Awareness of automated decision-making process.* • Algorithms are used to display multimedia content on platforms based on automated decisions. • Algorithms do not require human judgments in deciding which multimedia content to display on the platform. • Algorithms make automated decisions on which content I can see on the platform. *2. Awareness of risks and ethical issues.* • It is not always transparent why algorithms decide to display certain content. • The content recommended by algorithms on the platform may be subject to human biases and stereotypes. • Algorithms use personal data to recommend specific content on the platform, which affects online privacy. • Algorithms may exclude a user from seeing content other than their preference profile.	*1. Awareness of human-algorithm interaction.* • The content recommended by algorithms on the platform depends on users' online behavior on that platform. • The content recommended by algorithms on a platform depends on users' online behavioral data. • The content recommended by algorithms on a platform depends on users' available online data. *2. Development of bottom-up tactics.* • Optimizing web content to achieve a higher placement in the news feed. • Choosing to follow or unfollow certain accounts or hashtags to influence the composition of the news feed. • Modifying or creating content in specific ways to be favored by recommendation algorithms. • Exploiting feedback loops to reinforce specific patterns or outcomes.	*1. Positive, negative, indifferent emotional reactions.* • Feeling frustrated when not understanding why an algorithm shows certain content. • Feeling curious when encountering unexpected content suggested by algorithms. • Feeling indifferent about why algorithms select certain content. *2. Critical reflection generated by unexpected outcomes:* • Being prompted to reflect on algorithmic logics when the process suggests unexpected content. • Tendency to ignore unexpected outcomes of algorithm recommendations without seeking to understand the reason.

3. Conclusion

The paper aimed to illustrate the operational phase of the methodological framework at the first level of the ALGOFEED study, highlighting the importance of developing appropriate research tools to investigate the complex dynamics of interaction between users of digital platforms and

recommendation algorithms. The developed survey represents a significant step towards a thorough understanding of how users' technical and digital skills and platform usage practices influence and are influenced by algorithmic recommendations. A fundamental aspect emphasized in the paper is the importance of identifying and selecting representative user profiles that can provide a comprehensive overview of various interaction experiences with recommendation algorithms. This approach enabled us to grasp the complexity of the relationships between users and algorithms and analyze the various dimensions involved in detail. The empirical path useful for achieving this goal consists of creating synthetic indices, each corresponding to ideal-typical user profiles. This process is facilitated through analysis techniques aimed at developing models based on the combination of indicators selected during the operational data collection phase through interrogation, such as PLS Path Modeling. Therefore, the objective is to establish a typology categorizing online users types and magaing their exposure to feedback-loop processes. This pathway becomes viable upon integration with the remaining phases of the research. Subsequently, during the tracking and the interview stages, insights into the cultural consumption patterns of these user types emerge, facilitated by an understanding of their reactions to recommendations and their level of algorithmic awareness, obtained through the application of regression and clustering analysis. Within the scope of the initial research phase, the objective is to formulate three synthetic indices through the combination of indicators of dimensions related to digital skills and PUT. By combining these indicators (Figure 3), the synthesis of the audience-oriented user is postulated. This profile represents a user with medium-low content processing capabilities engaging on digital platforms as part of a diverse audience seeking entertainment content. Conversely, at the opposite end of the spectrum, we find the creator user - a user with high content processing capabilities carefully produced and promoted for professional and career advancement purposes in the digital creation sector. In the middle lies the uncertain user, straddling the characteristics of the two previous types, with a medium level of content processing capabilities and currently in a phase of consideration regarding a potential professional investment in the sector.

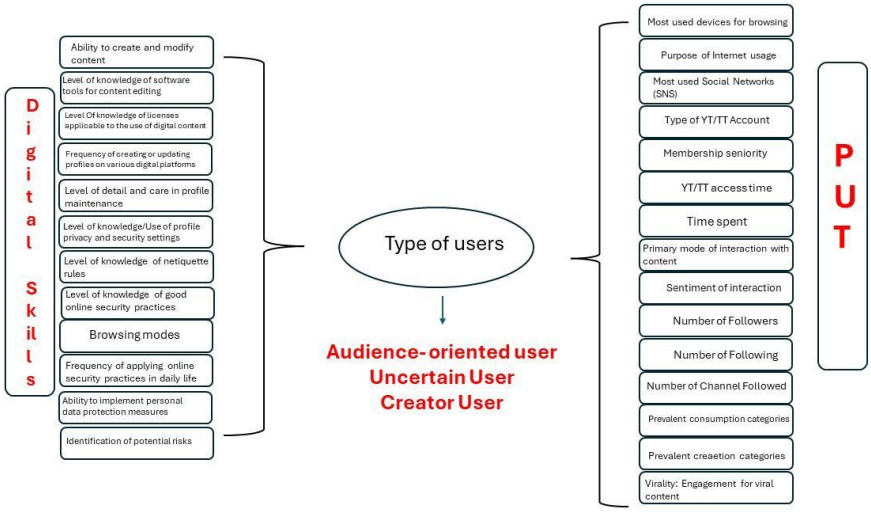


Figure 3 – The typological scheme obtained by the PLS-P

The crucial part of this paper was methodological: the operational definition of dimensions as digital skills, algorithmic awareness, and cultural consumption on social platforms that required a rigorous review of existing literature tools and careful methodological reflection. The project's subject matter contains some unexplored knowledge spaces (e.g., digital skills related to social media usage), making it essential to consider new dimensions and indicators for profiling our objectives.

The study broke down platform usage into the contexts of social platforms, subscription methods, and user experiences. It emphasized the necessity of considering users' platform presence, digital navigation skills, and familiarity with social media. Algorithmic awareness was identified as a critical area, highlighting its complexity and the nascent state of research tools for operationalization. The analysis covered cognitive, procedural, behavioral, and emotional aspects of users' evaluation and response to algorithmic recommendations in digital environments.

In conclusion, while it is possible to say that, in the context of our project, it is still too early to assess the tool's actual reliability and the validity of the concepts and indicators defined, this exploratory proposal, which aims to open debates that would allow for reflection aimed at improving the presented measurement tool, remains of decisive importance.

References

Airoldi M. (2021). The Techno-Social Reproduction of Taste Boundaries on Digital Platforms: The Case of Music on YouTube. Poetics, 89.

Ananny M., Crawford, K. (2018). Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability. New Media & Society, 20(3), 973–89.

Bennato, D. (2008). Le professioni del web, in Marco Pedroni e Paolo Volontè (a cura), La creatività nelle professioni, Bozen-Bolzano University Press, Bolzano, pp.75-88, 75.

Boccia Artieri, G., & Marinelli, A. (2018). Introduzione: piattaforme, algoritmi, formati. Come sta evolvendo l'informazione online. Problemi dell'informazione, 43(3), 349-368.

Bruns A. (2019). Are Filter Bubbles Real? Polity.

Diakopoulos, N. (2019). Automating the news: How algorithms are rewriting the media. Harvard University Press.

Felaco, C. (2022). Lungo la scala di generalità: le dimensioni della consapevolezza algoritmica. Sociologia Italiana.

Fourcade M. & Johns F. (2020). Loops, ladders and links: the recursivity of social and machine learning. Theory and society, 49(5), 803-832.

Gran A.B., Booth P. & Bucher T. (2020). To Be or Not to Be Algorithm Aware: A Question of a New Digital Divide? Inf., Comm. & Society, 24(12), 1779-1796.

Hallinan B. & Striphas T. (2016). Recommended for You: The Netflix Prize and the Production of Algorithmic Culture. New Media & Society, 18(1), 117–37.

Jiang R. et al. (2019). Degenerate Feedback Loops in Recommender Systems. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 383–90.

Møller Hartley, J., Bengtsson, M., Schjøtt Hansen, A., & Sivertsen, M. F. (2021). Researching publics in datafied societies: Insights from four ap-proaches to the concept of 'publics' and a (hybrid) research agenda. New Media & Society, DOI: 14614448211021045

Ragnedda, M. (2018). Conceptualizing digital capital. Telematics and informatics, 35(8), 2366-2375.

Rogers R. (2017). Digital methods for cross-platform analysis. The SAGE Handbook of Social Media, 91-110

Spohr, D. (2017). Fake news and ideological polarization: Filter bubbles and selective exposure on social media. Business information review, 34(3), 150-160.

Teddlie C. & Tashakkori A. (2011). Mixed methods research. The Sage Handbook of Qualitative Research, 4, 285-300.

Wiener N. (1989). The Human Use of Human Beings: Cybernetics and Society. Free Association Books.

Zarouali, B., Boerman, S. C., & de Vreese, C. H. (2021). Is this recommended by an algorithm? The development and validation of the algorithmic media content awareness scale (AMCA-scale). Telematics and Informatics, 62, 101607.

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17838

Challenges in Upholding Human Autonomy through the Right to be Forgotten

Sadaf Zarrin, Irene Unceta Mendieta
Esade Business School, Spain

How to cite: Zarrin, S.; Unceta Mendieta, I. 2024. Challenges in Upholding Human Autonomy through the Right to be Forgotten. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024.https://doi.org/10.4995/CARMA2024.2024.17838

Abstract

The paper examines the difficulties and challenges in implementing the right to be forgotten, highlighting the importance of this right for individual autonomy and privacy. It explores the main obstacles to upholding this right from legal, ethical, practical, and technical viewpoints, providing a summary of the existing problems and making recommendations for potential solutions. To improve the applicability of human rights in the digital world, the research emphasizes the significance of public awareness, international collaboration, and improvements in machine unlearning solutions. In order to assist the effective application of the right to be forgotten, the paper ends with suggestions for future research. These ideas seek to achieve a balance between autonomy, the need for privacy, and the rapid development of technology in digital spaces.

Keywords: *The Right To Be Forgotten; GDPR; Machine Unlearning, Human Right, Autonomy.*

1. Introduction

Machine Learning (ML) opens up new paths for innovation and expansion across various fields, such as online retail, medical care, education, legal practices, and national defense (Wirtz, 2019). ML applications in industry demonstrate abilities to learn adaptively and solve problems, broadening their application from creating new products to independently managing corporate operations (Mann, 2016). As a result, these technologies are quickly becoming a fundamental part of our everyday lives, influencing not just how we obtain products and services, but also altering the way we collect and analyze information, make decisions, and, ultimately, limiting our freedom to exercise those choices (Yeomans, 2015). In essence, ML technologies considerably affect our individual autonomy.

From a philosophical standpoint, autonomy is defined as the freedom to decide and act, along with the chance and freedom to implement our decisions (Pirhonen, 2020). While there are

various interpretations of autonomy (Gumbis, 2008) (Mackenzie, 2014), they all agree on the fundamental principle of self-determination. The United Nations (UN) encapsulates this consensus with its definition: “Autonomy is the acknowledgment of a person's right to hold views, to make choices and to take actions based on personal values and beliefs” (Gumbis, 2008).

Preventing industrial ML from undermining human autonomy is crucial for effective regulatory oversight within the European regulatory framework (Nikolinakos, 2023). However, the importance of autonomy has been largely overlooked in the designing and application of industrial ML technologies (Subías-Beltrán, 2023). Although current legal frameworks, like those set by the General Data Protection Regulation (GDPR), offer potential, their full integration as standard practices has not yet been achieved. In fact, the practical application of specific measures mandated by these laws has not been fully applied in the deployment of industrial ML systems. An illustrative example is *'the right to be forgotten'*.

The development of *the right to be forgotten* within EU data protection law and its formal acknowledgment by the European Court of Justice (ECJ) represent a proactive response to the evolving challenges of personal data protection in an increasingly digital world. This recognition notably progressed through the pivotal Google Spain SL v. Agencia Española de Protección de Datos (AEPD) case, where the ECJ addressed a complaint involving the request for Google to remove links containing outdated personal information. The Court's decision clarified that search engines act as data controllers and are, therefore, subject to data protection laws. This landmark ruling underscored the necessity for individuals to have the ability to control their digital footprint, particularly concerning information that is no longer relevant or necessary. By ruling in favor of the "right to be forgotten," the ECJ set a significant legal precedent, leading to the explicit inclusion of this right in the General Data Protection Regulation (GDPR), thereby embedding individual privacy and data control at the core of the digital age's legal framework which gives the autonomy to individuals to ask for the deletion of any of their data at any time they desire (Peguera, 2015).

Therefore, *the right to be forgotten* encompasses two key aspects, granting EU citizens the authority to request the deletion of any of their personal data. The first part specifically concentrates on the aspect of personal data usage, allowing individuals to demand the removal of their information as their digital right when it is no longer necessary, or if they withdraw consent or challenge the legitimacy of the data's processing. The second part is that the regulation mandates the elimination of any links to, or copies of, this information as well. Here, the focus shifts towards the technical aspect, emphasizing the need for mechanisms of machine unlearning (MU). This is particularly crucial as ML models often memorize training data (Bourtoule 2021). This characteristic renders the models susceptible to privacy attacks wherein adversarial opponents aim to extract information about the training data points. Such scenarios significantly compromise user secrecy and privacy (Huang, 2011). Thus, addressing this

technical aspect is vital for ensuring the comprehensive protection of personal data under the GDPR.

However, there are significant concerns regarding implementation of the right to be forgotten and although it represents a significant advancement towards protecting individual privacy as human right, it is not free from flaws and weaknesses, which diminishes its ability to safeguard human right in the digital realm. In light of this situation, we aim to answer to this question in this paper: 'What are the primary limitations of the right to be forgotten when applied in legal enforcement, and what steps can be taken to address these limitations to enhance the right's efficiency and applicability in the digital era?' In our endeavor to find the answer to this query, we reviewed various academic sources and identified diverse perspectives across different papers. We classified the limitations of *the right to be forgotten* into four primary groups: ethical dilemmas, legal and regulatory challenges, operational problems, and technical issues and discussed them in section 2. In Section 3, we present several recommendations to tackle these issues. Finally, in Section 4, 'Conclusion and Future Work', we underline the importance of continued exploration and development.

While the significance of data privacy and the right to be forgotten continues to grow, there's a noticeable gap in research that comprehensively addresses the various challenges associated with implementing these concepts. Previous studies have tended to focus on only one aspect, overlooking the broader picture. This research holds critical importance as it addresses the intricate balance between technological advancement and fundamental human rights in the modern age, challenges in implementing the right to be forgotten, technical obstacles related to machine unlearning, and the legal and ethical implications of data privacy. By identifying the limitations of the enforcement of *the right to be forgotten* in both protecting personal data and its technical implementation, this study aims to shed light on the complexities and challenges that arise from implementing such a right within the digital landscape.

2. Challenges of the Right to be Forgotten

In this section, we aim to provide a comprehensive understanding of the current challenges of implementing *'the right to be forgotten'* and set the stage for discussing potential solutions in the subsequent section. Some experts view *the right to be forgotten* as a form of internet censorship, as it may make finding pertinent information, or articles related to a person challenging or even unfeasible (Lee, 2015). On the other hand, there are arguments that this right can exist harmoniously with the freedom of expression and information, if there is a clear demarcation of their boundaries and an effective balance between them (ANGELES, 2016).

Moreover, there are various issues in the technical part of its implementation. If we utilize the data in a machine learning model, addressing the need to forget relevant data after training the model involves tackling challenges such as stochasticity and incrementality in ML algorithm

training, and catastrophic unlearning (Nguyen, 2022). Machine unlearning (MU) is designed to address situations where a user requests the deletion of specific data. In such cases, the previously trained model must undergo retraining to produce a new model. This updated model should reflect the distribution as though the deleted data had never been part of the initial learning process (Zhang, 2023). While there have been many proposed MU models, they are typically expensive and complicated, requiring either full or partial retraining of the model (Bourtoule, 2021) or complex matrix inversions (Liu, 2023). Furthermore, even if these methods were to prove effective, ensuring they comply with the regulation demands involves a deeper analysis of the legislation concerning the right to be forgotten to facilitate its translation into practice. This process includes clearly defining the situations where the right applies, choosing suitable technological solutions for enabling data to be forgotten in these contexts and integrating these solutions into the operational framework of industrial ML systems—a set of tasks that continues to pose significant challenges.

In the following subsections, we go through different categories of the mentioned issues.

2.1. Legal and regularity

As discussed in the introduction section, *the right to be forgotten* is defined under the General Data Protection Regulation (GDPR) and applies to EU citizens. However, the application of these regulations outside European Union introduces complexities such as balancing the EU's desire to extend its data protection norms worldwide against the principles of international comity and the legal diversity inherent in different nations. This balance is particularly precarious when it intersects with the concept of digital sovereignty, where countries may view the enforcement of EU standards within their jurisdictions as a form of 'data protection imperialism'. Consequently, this global push for EU data protection standards, including *the right to be forgotte*n, might lead to legal conflicts and contradictory rulings across different jurisdictions, underscoring the global intricacies of human autonomy in the digital age (Fabbrini, 2020).

Another legal issue of the right to be forgotten, in the context of the European Union's regulatory framework, is the intricate balance between the enforcement of this right under the General Data Protection Regulation (GDPR) and the obligations arising from the Electronic Identification, Authentication and Trust Services (eIDAS) Regulation. eIDAS establishes a standardized system for electronic identification and trust services across the EU, enhancing security and facilitating digital transactions and services. While promoting digital efficiency and cross-border interactions, eIDAS intersects with GDPR principles, particularly when it comes to personal data processing inherent in electronic identification schemes. The legal challenge here lies in harmonizing the robust identity verification mechanisms mandated by eIDAS, which are essential for digital market integration, with the stringent privacy rights protected by the GDPR, including the right to be forgotten (Andraško, 2021).

2.2. Ethical issues

The first important issue within the European legal framework is its struggle to balance the right to be forgotten with the fundamental rights of freedom of expression and information, defining the boundaries between an individual's privacy rights through data erasure and the public's interest in information access. This dilemma is exacerbated by the need for a legal mechanism that can effectively determine when personal data should remain accessible and when it should be removed, considering varying contexts and the evolving nature of digital information. The absence of a clear, universally applicable legal standard complicates navigating these conflicting rights, leading to uncertainties in the enforcement and application of the right to be forgotten. This legal challenge impacts not only data subjects and controllers but also broader societal values such as transparency and accountability, making it a critical area for legal refinement and development (Kocharyan, 2021). Another important ethical issue in implementing the right to be forgotten is the lack of awareness among organizations and individuals about the GDPR and its provisions, including the right to be forgotten. This lack of awareness can lead to significant challenges in ensuring the effective application and enforcement of this right (Addis, 2018).

2.3. Operational issues

The right to be forgotten faces issues due to its vagueness, such as unclear legal definitions and varied interpretations across regions. This uncertainty, rooted in evolving EU case law without solid legislative guidance, makes its enforcement challenging. It complicates decisions for data controllers on erasure requests and weakens privacy protection (Kocharyan, 2021).

2.4. Technical Issues

While implementing the right to be forgotten with the help of MU algorithms, some major concerns arise which we mention some of them here. The first problem is the challenge of MU algorithms to handle large amounts of data deletion efficiently in big data. This challenge requires high adaptability algorithms which can process and "forget" significant volumes of data without compromising the model's accuracy or performance (Zhang, 2023). Moreover, measuring the influence of each data point on the learning process before implementing the unlearning algorithm is not fully possible. This is compounded by the computational complexity of influence functions and the challenge of adapting them for complex models like deep neural networks (DNNs) which makes it difficult to catch the change in accuracy before implementing the MU model (Koh, 2017). The incremental nature of training, where updates reflect all previous updates, making the impact of any single training point implicitly influence all subsequent model updates is another problem while implementing Machine Unlearning. This further complicates the unlearning process since the removal of any data point affects the entire training history (Bourtoule, 2021). The trade-off in MU algorithms is mentioned as a problem in implementing it: to achieve a top-performing unlearning or in other words, high 'forget' quality, we must give up a high level of efficiency or utility (Kurmanji, 2024).

These are some problems and issues of implementing the right to be forgotten in practice which shows the complexity and multifaceted nature of data deletion, including technical challenges in completely erasing data without harming the integrity of existing datasets, legal and regulatory ambiguities across different jurisdictions, and the potential for unintended consequences such as compromising the accuracy of machine learning models or infringing on the public's right to information. In the discussion section, we focus on some solutions to mitigate some of these challenges and provide some future research ideas for the scholars to help implement *the right to be forgotten*, as an important aspect of human rights, more efficiently.

3. Discussion

Addressing the challenges mentioned in the previous section requires a multidisciplinary approach and operational considerations. in this section, we outline potential steps to mitigate these challenges and enhance the applicability of the right to be forgotten.

Increasing public awareness about their rights on digital platforms and understanding the right to be forgotten is essential. Educational programs in schools and universities teaching future workforce can be an effective resource in this matter. Moreover, campaigns and training programs in organizations, especially in the data-heavy sectors can train individuals to understand their and others' rights in the digital world and how to exercise them.

All countries, especially countries active in the digital field, should work together to develop an international agreement to enhance human autonomy and reduce legal conflicts.

Machine Unlearning is a newly developed concept and there is a great deal of work that scholars can focus on to enhance efficiency in this area and address the technical challenges associated with data deletion. Focus of the computer and data science scholars to create secure, privacy-preserving, and scalable algorithms that facilitate the removal of data can help to implement the right to be forgotten more effectively without affecting the performance significantly.

A serious effort to refine the legal definitions related to the right to be forgotten is needed to reduce confusion and help more transparent implementation of laws in this area. The development of operational guidelines and best practices can help organizations manage the complexities of implementing the right to be forgotten. Establishing clear processes to assess the response to data deletion, monitoring, and auditing can be part of this process.

In Section 4, we provide some ideas for future work and emphasize the necessity for ongoing interdisciplinary collaboration among scholars to address the challenges of the right to be forgotten and machine unlearning.

4. Conclusion and Future Work

The focus of this paper is on *the right to be forgotten* in the digital era as an important human right in today's world and its challenges in legal, ethical, and technical aspects highlighting the complex interplay between autonomy, privacy, and digitalization. Since the paper discusses a crucial aspect of data privacy and security, it is relevant to web and big data practitioners who are responsible for managing and securing large volumes of personal data. There is a significant opportunity for computer and data scientists to research different techniques and algorithms of MU and data deletion techniques which would help to enhance the accuracy and efficiency of the models specifically in large-scale data. Moreover, research should be done to check the level of awareness of different groups of society about their rights and to plan awareness raising policies accordingly. Furthermore, it seems that with the high speed of artificial intelligence and the digital world, the creation of ethical guides and rules in this world is far behind. Consequently, creating a framework for guiding and deploying artificial intelligence systems concerning autonomy, privacy, and human rights is a necessity.

In conclusion, enhancing the right to be forgotten requires a concerted international cooperation of policymakers, technologists, and legal experts. By addressing the identified challenges in this research through legal reforms, public engagement, and technical innovation, we can become closer to respecting privacy and human autonomy. Future research in this relatively new area will play a significant role in the evolution of privacy rights and its alignment with rapid technological change.

References

Addis, M. C. (2018). he general data protection regulation (GDPR), emerging technologies and UK organisations: awareness, implementation and readiness.

Andraško, J. &. (2021). Those who shall be identified: The data protection aspects of the legal framework for electronic identification in the European Union. *TalTech Journal of European Studies, 11*(2), 3-24.

Angeles, L. O. (2016). Entertainment Law Review.

Bourtoule, L. &.-C. (2021). Machine unlearning. *IEEE Symposium on Security and Privacy (SP)*, (pp. 141-159).

Bourtoule, L. C.-C. (2021). Machine unlearning. *In 2021 IEEE Symposium on Security and Privacy (SP)*, (pp. 141-159).

Fabbrini, F. &. (2020). The right to be forgotten in the digital age: The challenges of data protection beyond borders. *German law journal, 21*, 55-65.

Gumbis, J. B. (2008). Do human rights guarantee autonomy? *Cuadernos Constitucionales de la Cátedra Fadrique Furió Ceriol, 62*, pp. 77-93.

Huang, L. J. (2011). *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence.*

Kocharyan, H. V. (2021). Critical views on the right to be forgotten after the entry into force of the GDPR: Is it able to effectively ensure our privacy? *International and Comparative Law Review*, *21*(2), 96-115.

Koh, P. W. (2017). Understanding black-box predictions via influence functions. *In International conference on machine learning*, (pp. 1885-1894).

Kurmanji, M. T. (2024). Towards unbounded machine unlearning.

Lee, E. (2015). The right to be forgotten v. free speech. *ISJLP*, *12*, 85.

Liu, J. X. (2023). Muter: Machine unlearning on adversarially trained models. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, (pp. 4892-4902).

Mackenzie, C. (2014). Three dimensions of autonomy: A relational analysis. In Autonomy, oppression and gender. 15-41.

Mann, G. &. (2016). Hiring algorithms are not neutral. *Harvard Business Review*.

Nguyen, T. T. (2022). A survey of machine unlearning. *arXiv preprint arXiv:2209.02299.*

Nikolinakos, N. T. (2023). EU Policy and Legal Framework for Artificial Intelligence, Robotics and Related Technologies-The AI Act. *Springer*.

Peguera, M. (2015). In the aftermath of Google Spain: how the 'right to be forgotten'is being shaped in Spain by courts and the Data Protection Authority. *International Journal of Law and Information Technology*, *23*(4), 325-347.

Pirhonen, J. M. (2020). Could robots strengthen the sense of autonomy of older people residing in assisted living facilities?—A future-oriented study. *Ethics and Information Technology*, *22*(2), 151-162.

Subías-Beltrán, P. P. (2023). Respect for Autonomy in the Machine Learning Pipeline. In Artificial Intelligence Research and Development. *VIOS Press*, 221-230.

Wirtz, B. W. (2019). . Artificial intelligence and the public sector—applications and challenges. *International Journal of Public Administration*, *42(7)*, pp. 596-615.

Yeomans, M. (2015). What every manager should know about machine learning. *Harvard Business Review*, *93(7)*.

Zhang, H. N. (2023). A review on machine unlearning. *SN Computer Science*, *4*(4), 337.

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17839

Towards Intangible Value Quantification: Scope, Limits & Shortages of Artificial Intelligence application

Salvador Domínguez Gil[1], Andrea San José Cabrero[2], Antonio Sánchez Gea[3], Pilar Miguel-Sin[1], Gema Ramírez Pacheco[1]

[1]Universidad Politécnica de Madrid, Spain, [2]Universidad de Navarra, Spain, [3]SOCOTEC, Spain

How to cite: Domínguez Gil, S.; San José Cabrero, A.; Sánchez Gea, A.; Miguel-Sin, P.; Ramírez Pacheco, G. 2024. Towards Intangible Value Quantification: Scope, Limits & Shortages of Artificial Intelligence application. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17839

Abstract

The application of Artificial Intelligence (AI) in the realm of economic markets, particularly in the business and real estate sectors, has witnessed substantial growth. However, its effectiveness is curtailed by several limitations, especially in the context of the rising valuation of intangible assets. The intangible nature of assets such as brand value, environmental impact or social impact, among others, presents a challenge for AI, which relies on quantifiable data for analysis and decision-making. The intrinsic volatility and uncertainty of markets, heightened by the intangible asset valuation, further complicate the AI's predictive accuracy and adaptability.

AI models, primarily dependent on historical data, struggle to accurately forecast market movements influenced by intangible factors, which are often subjective and dynamically changing. This limitation is particularly pronounced in the real estate promotion sector, where the perceived value of properties can be significantly affected by intangible elements like location prestige or architectural uniqueness. Additionally, the ethical implications of AI deployment, such as data privacy concerns and potential biases in algorithmic decision-making, pose further constraints on its application in these sectors.

While AI offers transformative potential for economic markets, its current limitations in handling the valuation of intangibles, market volatility, and ethical considerations necessitate a cautious and complementary approach to its integration into business and real estate promotion strategies, specially in the concern of life-cycle approaches.

Keywords: *Real Estate Assessment; Life-Cycle Assessment; Intangible Value; Sustainable Development; Data-driven decision making.*

1. Introduction

In the contemporary era, the quantification of intangible values within society poses a formidable challenge in the fields of social and economic sciences. These values, encompassing aspects such as culture, identity, social well-being, and community cohesion, among others, play a pivotal role in the sustainable development of societies. However, their intangible nature and the inherent complexity in measuring them present significant methodological and conceptual hurdles.

From a scientific and professional perspective, it is crucial to address the quantification of these values through approaches that capture their essence and impact on social, economic and environmental aspects. Traditionally, the measurement of progress and development in cities has focused on tangible and quantitative indicators, such as Gross Domestic Product (GDP). While these are essential, they do not encompass the full spectrum of what constitutes the well-being and quality of life for individuals. In urban development and constructions, the short-term economic perspective has shifted towards life-cycle asset managemenet approaches, largely due to the emergence of sustainable financing mechanisms based on new sustainability indicators and life-cycle approaches. Tus, the objective linked to the valuation of intangibles by AI based systems is deeply linked to improving stakeholders decision-making processes and defining the basis for common assessment methodologies.

To this end, various qualitative and mixed methodologies have been developed to integrate the valuation of intangible into socioeconomic analysis, though few have achieved significant international depth. These methodologies include perception surveys, ethnographic studies, discourse analysis, and social, economic or environmental impact assessments, offering a more holistic and representative approach to social reality. However, as stated by many authors, their subjective and qualitative nature presents challenges for Artificial Intelligence algorithms, bringing large gaps between information and decision-making processes.

2. Limitations of Artificial Intelligence in intangible valuation

The management and quality of data sources for the quantification of intangible values face significant challenges even for neural network models. The diversity and subjectivity inherent in these values make standardizing data collection and analysis methods difficult. Additionally, the lack of comprehensive and reliable databases hampers the ability to perform robust temporal and spatial comparisons, compounded by variability in the quality of collected information, which may be influenced by cultural, political, and social biases.

On the first hand, the scarcity of data represents a significant obstacle in the effective implementation of AI in the real estate sector. This point is common to Life Cycle Assessment (LCA) methodologies (Schneider-Marin & Lang, 2020) that are to be applied by AI. Artificial

Intelligence models require large datasets to be trained and generate accurate predictions, which will depend on the type of intelligence applied and the selected model. However, in the case of unique real estate assets or in less developed markets, the availability of historical and current data may be limited. This data scarcity can lead to AI models that are not adequately informed, resulting in less reliable predictions and evaluations.

In the field of Life Cycle Cost Assessment in buildings, there are "elementary" concepts that can be measured and identified in data-bases, as costs of construction, operation, maintenance and demolition of buildings. On the other side, the integral costs or Whole Life Costs include these as well as externalities, indirect costs, intangible costs, economic benefits, rents or revenues, as well as environmental or social costs. Although some studies analyse and develop several of the phases linked to both concepts, many of these aspects are difficult to asses due to the lack of data or methodologies. In fact, there are no empirical studies in the last ten years that validate, with verifiable data, all or part of their theoretical determinations. That may be due to the sensitive nature of the information flows of supply markets, of the agents in the construction and real estate sectors, as well as to the uncertainty regarding the evolution of future costs and revenues linked to the development of markets and society itself (Enshassi et al., 2014)

The concept of externality, the direct and indirect effects that an activity has on other activities, connects in this contexts with the concept of intangible and, at the same time, with the concepts of the "elementary" costs stated previously. Intangibles costs, known as all the costs and revenues that can be foreseen and assumed but do not occur until a new practice or policy is implemented, can be valuated under the Monte Carlo methods, implemented in AI systems. However, the scope and scenario definition, according to ISO 15686 and other standards, may bring further gaps in AI applications due to the need of continuous external expert judgement.

Additionally, intangible impacts can be seen as direct economic impacts on the customer's organization, in terms of asset management, which result in improvements for users well-being that can bring further economic implications (ISO 15686-5:2017, 2017). Such is the case of concepts liked to functionality, flexibility, or even aesthetics, which can influence the evaluation, which play a major role in the investment-profit puzzle (Orhangazi, 2019).

The evaluation and interpretation of results derived from the measurement of these intangible values also present considerable challenges. The qualitative nature of much of this data necessitates an interpretive approach, which, while enriching the analysis, can complicate the communication and understanding of findings. In this context, developing theoretical and methodological frameworks that facilitate the explanation and interpretation of results, as well as their integration into policy and economic decision-making, is crucial.

3. Intangible values in Smart Cities and urban development tendencies

In the context of future smart cities, the quantification of intangible values gains particular relevance. These cities, characterized by the integration of information and communication technologies into urban management, aim to enhance the quality of life of their inhabitants and promote sustainable development. Measuring intangible values in this context allows for the assessment of the impact of implemented policies and technologies on aspects such as social cohesion, citizen participation, cultural identity, and environmental sustainability. Therefore, incorporating indicators related to these values in the development, planning, and management systems of smart cities is essential to ensure that technological development translates into tangible benefits for the community.

AI encounters substancial challenges in comprehending urban development, fashion trends, social value shifts and their consequential impacts on the valuation of cities, spaces and buildings. The inherent complexity of these phenomena arises as intangibles from their deeply contextual, dynamic and multifaceted nature, which is significantly influenced by human behaviors, cultural nuances and societal norms. These concepts of difficult quantification often elude algorithmic predictability.

Urban development encapsulates a myriad of variables including economic, environmental and social dimensions that are intelinked in intricate ways, rendering simplistic computational models inadequate. Fashion trends, reflecting ephemeral societal preferences and cultural expressions, add another layer of volatility that AI systems struggle to accurately forecast or interpret due to their rapidly evolving nature. Furthermore, social values, which underpin the collective perception and importance of various aspects of urban life, are subject to continuous transformation influenced by global events, technological advancements and shifting cultural paradigms.

4. Conclusions

The quantification of intangible values in society presents a set of methodological, technical and conceptual challenges that require an interdisciplinary and collaborative approach, extending beyond the scientific-mathematical framework. Overcoming these difficulties will not only allow for a better understanding of the social development and its dynamics but will also facilitate the formulation of more effective public policies, aligned with the well-being and aspirations of the community in the matter of sustainable development.

In this endeavor, methodological innovation, improved data managemenet and the promotion of a culture of continuous evaluation will be key elements in advancing towards a more inclusive, resilient and sustainable society. Thus, the relationship between objectives and decision-making tools should also be revisited to ensure alignment and effectiveness.

This variability challenges AI's ability to adapt and remain relevant in real-time analysis and decision-making processes, thereby limiting its efficacy in providing actionable insights for urban planning, real estate valuation and design of socio-economic development strategies.

References

Bahramian, M. & Yetilmezsoy, K. (2020).Life cycle assessment of the building industry: An overview of two decades of research (1995–2018). Energy and Buildings, 219, 109917. https://doi.org/10.1016/j.enbuild.2020.109917

Enshassi, A., Kochendoerfer, B. & Rizq, E. (2014). Evaluación de los impactos medioambientales de los proyectos de construcción. Revista ingeniería de construcción, 29(3), 234-254. https://dx.doi.org/10.4067/S0718-50732014000300002

ISO. (2017). ISO/DIS 15686-5:2017, Building and constructed assets — Service life planning — Part 5: Life Cycle Costing.

Orhangazi, Ö. (2019). The role of intangible assets in explaining the investment–profit puzzle. In Cambridge Journal of Economics (Vol. 43, Issue 5). https://doi.org/10.1093/cje/bey046

Schneider-Marin, P. &Lang, W. (2020). Environmental costs of buildings: monetary valuation of ecological indicators for the building industry. International Journal of Life Cycle Assessment, 25(9), 1637–1659. https://doi.org/10.1007/s11367-020-01784-y

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17455

Digitalisation - the Basis for Building an Agile Enterprise

Andrea Janáková Sujová, Petra Lesníková
Department of Economics, Management and Business, Technical University in Zvolen, Slovakia

How to cite: Janáková Sujová A.; Lesníková P. 2024. Digitalisation - the Basis for Building an Agile Enterprise. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17455

Abstract

Enterprise digitalisation and agility are two key concepts that can work together to contribute to an organisation's competitiveness. Enterprise digitalization as a process of transforming traditional business models and processes using digital technologies enables the implementation of agile principles. Enterprise agility as the ability to adapt quickly and efficiently to unpredictable changes in the environment is becoming an important competitive factor. The aim of the article is to reveal the interactions between digitalization and enterprise agility and to present the results of primary research in industrial enterprises of the Slovak Republic focused on the perception of the importance of digitalization in the context of agility. The results showed that digitalization is an important element of agility and an essential starting point in building an agile enterprise. Slovak industrial enterprises consider digitisation as an important help in coping with unexpected changes such as the coronacrisis, as a result of which digitisation has accelerated. However, the current adverse global circumstances mean that one third of enterprises have reduced or stopped digitisation altogether and the number of digitising enterprises has declined over three years. Digitisation can transform business models and create agile operating models.

Keywords: *digital transformation; enterprise agility; rapid adaptation; industrial enterprises; agile business model*

1. Introduction

IT-enabled business transformation came to the fore with the commoditization of computer technology and the spread of the Internet in the 1990s and has recently revived in the wake of global crises such as the coronacrisis and the energy crisis. The adoption of digital technologies affects almost all areas of firms, such as production, organizational structures, and relationships with partners. The use of digital technologies to create new or modify existing business models and processes or to support the transformation of organizational structures, resources or relationships with internal and external stakeholders is also referred to as digital transformation

(DT) (Vial, 2019). Plekhanov et al. (2023) outlined a layered model of DT with three layers: organizational core, organizational periphery, and external environment. Regardless of whether DT-induced changes are bottom-up or top-down, firms eventually evolve into interconnected networks of decentralized communication channels that do not follow the rules of traditional vertical hierarchies. Digital technologies have the potential to improve resource efficiency, optimize production processes, and strengthen risk management (Kusiak, 2018). Digital business models are characterised by accelerated rates of value creation and changes in resource management (Paiola and Gebauer, 2020). Digitalization of business models tends to trigger subsequent innovation, contributing to a chain reaction of interrelated and co-dependent innovation activities (Wiesbock & Hess, 2019). Many industries are moving towards shorter innovation cycles due to advances in digital simulations, reduced product creation costs and shorter time-to-market requirements (Rossi et al., 2020). According to Kwilinski (2023), there is a link between sustainable development and digital transformation; the introduction of digital tools, new methods and data processing methods is a priority for the development of the different components of sustainable development. Several sources conclude that DT has the potential to significantly increase the sustainability of business operations by enabling automation, designing smart solutions and facilitating direct communication between producers and customers (Sklyar et al., 2019).

In order to enable the effective adoption and use of digital technologies, firms are changing their organisational structures and establishing innovation labs, corporate innovation centres and so-called digital business units, which have greater autonomy and dedicated budgets (Seran & Bez, 2021). Autonomous teams contribute to higher organisational agility, shorter innovation cycles and flexibility. Enterprise agility is a firm's ability to proactively respond to changing customer demands and market trends by adapting and reconfiguring organizational processes and the delivery of products and services (Brock & Von Wangenheim, 2019). Agile approaches based on frequent and rapid experimentation can improve firms' responsiveness to technological change and competitive pressures. Becoming an agile firm with the ability to respond quickly and effectively to changes in the global business environment is a necessity for business. The role of IT/IS in the context of enterprise agility is undeniable, IT helps to ensure agility mainly by accelerating decision making and effective communication (Kocu, 2018). In the context of DT, agility means learning from failure and increasing the speed of development of digital products and services, that requires appropriate structures that enable rapid adaptation and can take advantage of opportunities arising from digital technologies (Baiyere et al., 2020). Sjodin et al. (2020) pointed out that agile and short process cycles are essential to accelerate innovation that is always up-to-date with technological advances and customer preferences. Various approaches such as Scrum, autonomous cross-functional teams, and continuous feedback loops can be used to help firms achieve agility (Guinan et al., 2019). It is important to note that the ability of firms to adapt to the environment is mainly supported by adequate competencies. An

agile approach requires dynamic capabilities, which Teece et al. (2016) characterize as a firm's ability to innovate, adapt to change, and create changes that are favorable to customers and unfavorable to competitors.

The aim of the paper is to identify the interactions between digitalization and enterprise agility and the perception of the importance of digitalization in the context of agility by managers of Slovak industrial enterprises. The main hypothesis was stated: Is digitalization leading enterprises to become agile?

2. Methodology

The first part of the research focuses on the analysis of existing knowledge in the areas of digitalization and enterprise agility in order to identify and summarize aspects of the relationship between these two concepts (phenomena). The second part of the paper presents the results of a primary questionnaire survey in Slovak industries conducted by Trexima. The aim of the survey was to determine the perception of the importance of digitalization in the context of coping with unexpected changes caused by the coronacrisis, which required the application of an agile approach and tested the ability of agility in enterprises. The research sample consisted of 57 respondents from the engineering, automotive and electrical engineering industries. The questionnaire survey was conducted repeatedly, between 2000 and 2022, to capture trends and track developments. More than 50% of the research sample consisted of large enterprises with over 249 employees and foreign ownership. Descriptive statistics methods and time series trend analysis were used to evaluate the results.

3. Results

3.1. Identifying the interactions between agility and enterprise digitalisation

From the analysis and summarization of theoretical and scientific knowledge, it is possible to identify the interactions between digitalization and enterprise agility in several aspects, as shown in Table 1.

Table 1. Relationships between agility and enterprise digitalization. Source: own.

Aspect	The principle of agility	A feature of digitisation	Interaction
Ability to adapt quickly	Agile methods are based on the ability to react quickly to changes	Allows you to better adapt to changes and technological trends	Digitisation provides the means to implement change quickly
Improving process efficiency	Flexible and efficient processes are the key to agility	Enables automation and optimization of processes	Digitisation simplifies and speeds up processes
Improved communication and cooperation	Fast and open communication	Enables effective information exchange and cooperation	Digital tools support agile working methods
Faster innovation	Rapid testing of new ideas	Supports innovation processes, enables the creation of new products and services	Digitalisation provides the means to introduce innovations quickly.
Increasing competitiveness	Rapid adaptation to changes in the competitive environment	DT enables faster and more efficient delivery of value to customers	Digitalisation makes delivery and meeting customer requirements faster and efficient

As can be seen from the data in Table 1, digitalisation and enterprise agility are intrinsically linked, with digitalisation forming the basis for the application of agile principles.

3.2. Perception of the importance of digitalization in industrial enterprises in the Slovak Republic

23% of industrial enterprises in the Slovak Republic are undergoing digital transformation today, despite the fact that up to 85% of enterprises that have not yet started to change see the need for digital transformation. However, in a three-year view of the share of digitizing firms, we see a continuous decline (2020 - 35%, 2021 - 26%, 2022 - 23%) (Industry 4UM, 2022). From the survey results, we have extracted the findings regarding the attitudes of businesses towards digitalisation in conjunction with unexpected changes due to the impact of the coronacrisis, as presented in Figures 1-4. The data in the graph in Figure 1 shows that businesses recognise the importance of digitisation in conjunction with a corona crisis, with 52% confirming that more significant digitisation would help to better prepare them to cope with changes during a corona crisis.

The graph in Figure 2 presents the impact of the corona crisis on the approach to digitalisation in the enterprise. Of the enterprises that had already started digitisation, 58% continued to make transformational changes in 2020 as they did before the pandemic, but this proportion dropped

to 50% within three years. The proportion of enterprises that have reduced (from 16% to 25%) or stopped (from 8% to 10%) digitization has gradually increased. 11% of enterprises are implementing even more intensively, and the number of enterprises that plan to resume digitization processes soon has dropped from 9% to 4%.

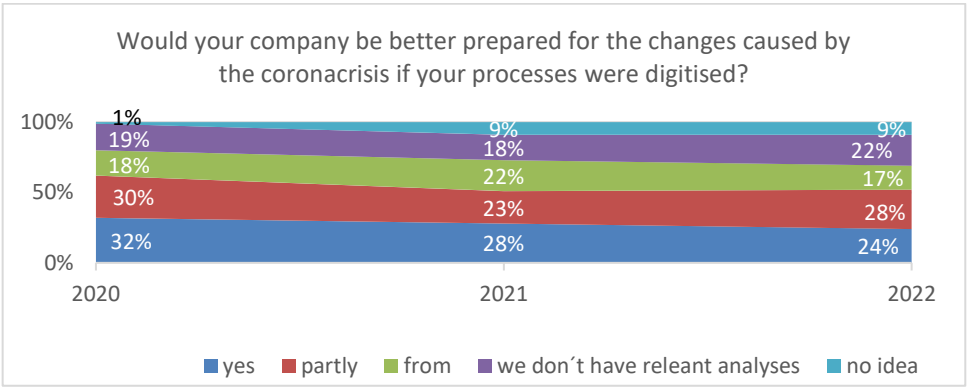


Figure 1: The importance of digitalization for corona crisis preparedness. Source: own.

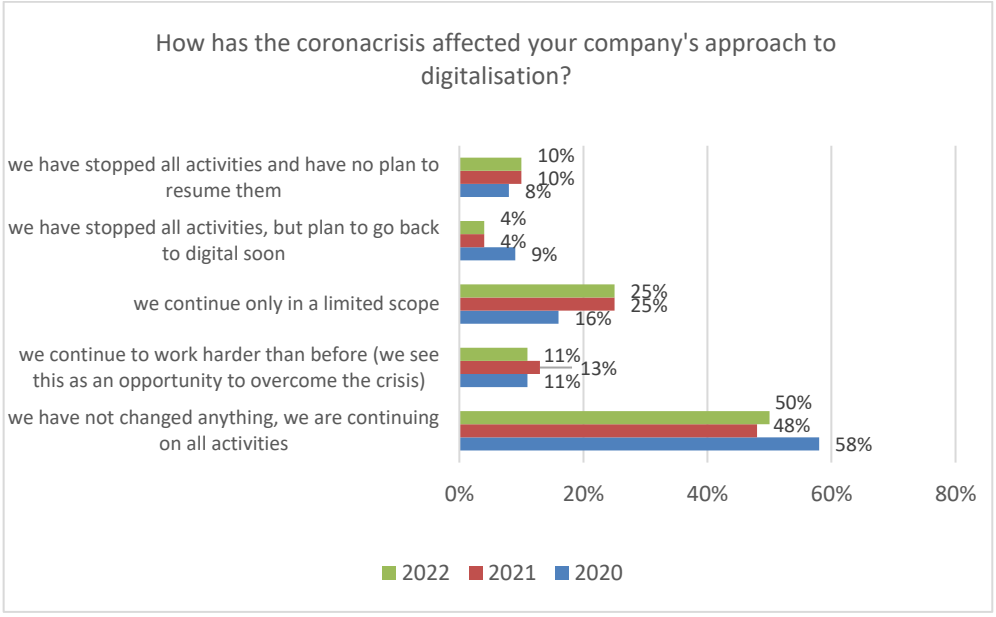


Figure 2. Businesses' approach to digitalisation during the corona crisis. Source: own.

The impact of corona crisis on the change in the rate of DT implementation is shown in Figure 3. Three thirds of enterprises did not know or did not assess this impact, but 22% said that the changes caused by the corona crisis had accelerated the digitisation processes in enterprises, while in 10% it had slowed them down.

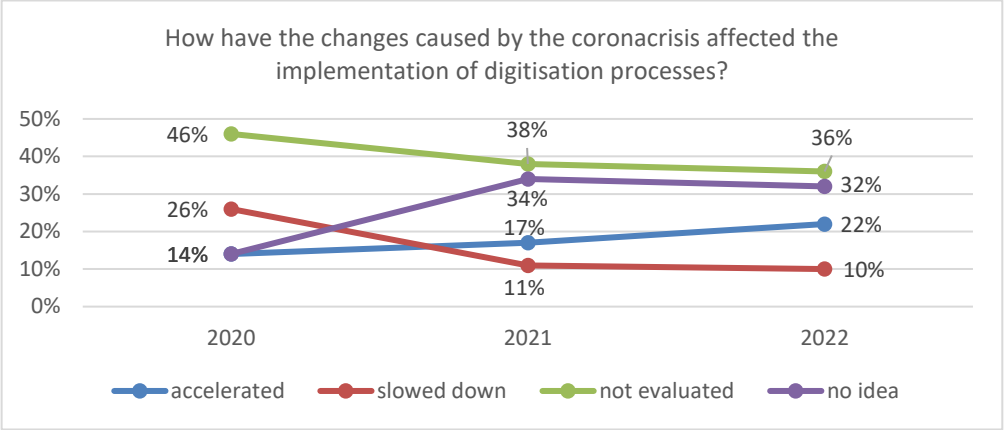


Figure 3: Impact of coronary crisis on the implementation of digitisation processes. Source: own.

When asked whether digitisation can change the business model of an enterprise, over 60% of enterprises said yes, with the proportion of enterprises saying yes rising steadily over the three years from 61% in 2020 to 68% in 2022. On average, 20% said no, and the number of enterprises that could not say gradually fell from 19% in 2020 to 14% in 2022. The results are shown in the graph in Figure 4.

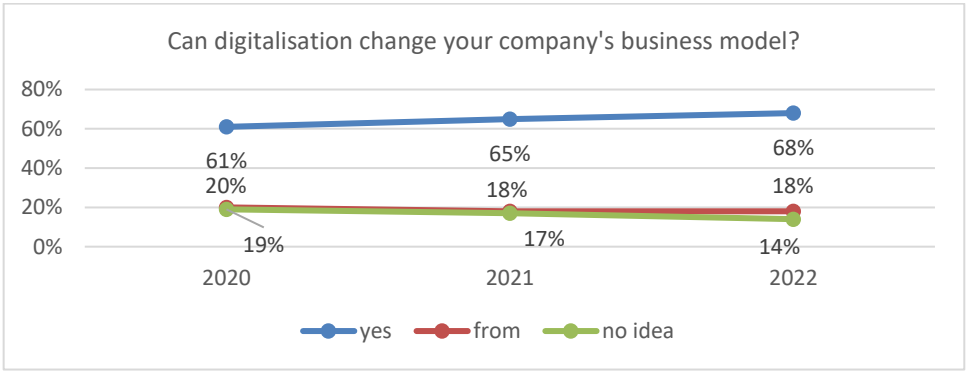


Figure 4. The opportunity to change the business model by going digital. Source: own.

However, according to more detailed findings, only foreign-owned companies are digitising almost exclusively, mostly applying the know-how of their parent companies. On the contrary, the share of Slovak companies in the monitored parameters is appallingly low and critically stagnant in terms of development. Companies see the need to digitise primarily in production processes and logistics, and in cooperation with customers, suppliers and buyers. One of the key challenges that the survey repeatedly identifies is building a corporate culture that supports the digital transformation of businesses. Enterprises lack strategic and application teams, do not build an environment for innovation development, and do not properly understand the role of

management in relation to the preparation and management of digital strategy. A fundamental barrier to the development of digitalisation is the lack of knowledge about digitalisation in industry. Businesses want to digitise but do not know how. More than half of enterprises lack the information they need to apply digital solutions and do not know where to get it.

4. Discussion and Conclusions

Digitising the business brings a number of benefits, including faster decision-making, increased efficiency and accuracy, improved competitiveness, innovation and better meeting customer needs. Organisations that successfully integrate digital technologies into their business are more likely to become agile and thrive in a rapidly changing business environment. Digital technologies enable faster responses and adaptation to change by providing the necessary resources and tools. In the difficult situation of the coronacrisis, industrial enterprises in the Slovak Republic have become more aware of the importance of digitisation for current stabilisation, future development and maintaining competitiveness, as the results of the survey presented here show. However, the pace of transformation is hampered by rising inflation as well as uncertainty linked to domestic political turmoil and the war conflict in Ukraine. These circumstances are setting priorities for businesses and influencing attitudes towards digitalisation. Transformational change has been halted or limited by 34% of enterprises due to adverse global circumstances. Barriers to the adoption of digital technologies are also barriers to the implementation of agility elements and thus to building enterprise agility. According to the results of the survey in Slovakia, the barriers of Slovak industrial enterprises include mainly lack of information, digital literacy of employees, lack of management support and corporate culture.

Enterprise digitisation and agility are interlinked and can be mutually reinforcing. Organisations that successfully combine these two aspects are more likely to achieve sustained success in today's dynamic and competitive business environment. Based on our findings, it can be concluded that digitalisation is an important starting point for building enterprise agility. The main message for practitioners is that going digital will enable the enterprise to be agile.

Acknowledgement

The paper is a partial result of the grant scientific project VEGA 1/0333/22.

References

Baiyere, A., Salmela, H., & Tapanainen, T. (2020). Digital transformation and the new logics of business process management. European Journal of Information Systems, 29(3), 238–259. https://doi.org/10.1080/0960085X.2020.1718007

Brock, J. K. U., & Von Wangenheim, F. (2019). Demystifying AI: What digital transformation leaders can teach you about realistic artificial intelligence. California Management Review, 61(4), 110–134. https://doi.org/10.1177/1536504219865226

Guinan, P. J., Parise, S., & Langowitz, N. (2019). Creating an innovative digital project team: Levers to enable digital transformation. Business Horizons, 62(6), 717–727. https://doi.org/10.1016/j.bushor.2019.07.005

Koçu, L. (2018). Business-IT alignment effects on business agility. International Journal of Commerce and Finance, 4(2), 60-93.

Kusiak, A. (2018). Smart manufacturing. International Journal of Production Research, 56(1-2), 508-517. https://doi.org/10.1080/00207543.2017.1351644

Kwilinski, A. (2023). The relationship between sustainable development and digital transformation: bibliometric analysis. Virtual Economics, 6(3), 56-69 https://doi.org/10.34021/ve.2023.06.03(4)

Paiola, M., & Gebauer, H. (2020). Internet of things technologies, digital servitization and business model innovation in BtoB manufacturing firms. Industrial Marketing Management, 89, 245–264. https://doi.org/10.1016/j.indmarman.2020.03.009

Plekhanov, D., Franke, H., & Torbjorn, H., N. (2023). Digital transformation: A review and research agenda. European Management Journal, 41(2023), 821–844

Rossi, M., Festa, G., Devalle, A., & Mueller, J. (2020). When corporations get disruptive, the disruptive get corporate: Financing disruptive technologies through corporate venture capital. Journal of Business Research, 118, 378–388. https://doi.org/10.1016/j.jbusres.2020.07.004

Seran, T., & Bez, S. M. (2021). Open innovation's "multiunit back-end problem": How corporations can overcome business unit rivalry. California Management Review, 63(2), 135–157. https://doi.org/10.1177/0008125620968609

Sjodin, D., Parida, V., Kohtamäki, M., & Wincent, J. (2020). An agile co-creation process for digital servitization: A micro-service innovation approach. Journal of Business Research, 112, 478–491. https://doi.org/10.1016/j.jbusres.2020.01.009

Sklyar, A., Kowalkowski, C., Tronvoll, B., & Sorhammar, D. (2019). Organizing for digital servitization: A service ecosystem perspective. Journal of Business Research, 104, 450–460. https://doi.org/10.1016/j.jbusres.2019.02.012

Teece, D., Peteraf, M. & Leih, S. (2016). Dynamic capabilities and organizational agility: Risk, uncertainty, and strategy in the innovation economy. California Management Review, 58(4): 13-35. http://dx.doi.org/10.2139/ssrn.2771245

Vial, G. (2019). Understanding digital transformation: A review and a research agenda. The Journal of Strategic Information Systems, 28(2), 118–144. https://doi.org/10.1016/j.jsis.2019.01.003

Wiesbock, F., & Hess, T. (2019). Digital innovations: Embedding in organizations. Electronic Markets, 30(1), 75–86. https://doi.org/10.1007/s12525-019-00364-9

https://industry4um.sk/prieskum-industry-4-0/

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17500

Work Realities and Behavioral Risk Factors in Italy

Angela Andreella, Stefano Campostrini
Department of Economics, Ca' Foscari University of Venice, Italy.

How to cite: Andreella, A.; Campostrini, S. 2024. Work Realities and Behavioral Risk Factors in Italy. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17500

Abstract

The connection between health, work environment, and job characteristics is a relevant issue in public health. However, it is often underexplored due to a lack of reliable data. To address this gap, we have delved into the subject using data from an NCDs-risk factor surveillance system (PASSI). We have examined information collected from respondents regarding their occupations relating to risk factors and health status. The proposed analysis employs text mining and cluster approach for categorical variables to identify sub-populations characterized by different socio-economic situations, risk factors, and job types. Although further analyses are needed to explore the potential of this approach better, initial results are promising. They highlight the practical implications of our findings for public health policies. For example, we found that occupations related to the building industry (for males) and healthcare professions (for females) appear to be associated with higher behavioral risk factors, which could inform targeted interventions.

Keywords: *Job, behavioral risk factors, surveillance system PASSI, categorical data, text mining, clustering*

1. Introduction

Behavioral risk factors, such as smoking, inadequate nutrition, excessive alcohol consumption, and physical inactivity, collectively referred to as SNAP, constitute significant contributors to morbidity and mortality. These factors are prevalent in high-income countries and increasingly in lower-income nations (Noble et al., 2015). Their impact extends to developing chronic diseases, the leading causes of global mortality (World Health Organisation, 2014). Recognizing how risk factors coalesce informs effective preventive interventions. This insight targets specific populations for tailored health promotion and emphasizes the potential impact of addressing multiple risk behaviors concurrently, a key to more impactful public health outcomes (Prochaska & Prochaska, 2011).

It is well known in the literature that SNAP risk factors correlate with other aspects of life, including socio-economic status, such as educational level and economic situation (Flaskerud et al., 2012; Minardi et al., 2011). This, in turn, also leads to a relationship between SNAP and the type of job and work environment, acknowledging the profound impact of work characteristics on health. However, the connection between job characteristics and health/SNAP remains underexplored mainly due to the lack of proper data. For that, a predominant focus on job stress as the connecting factor to behavioral risk factors persists in many analyses, often relying on investigations concentrated on specific occupations. Notable examples include the research of Kouvonen et al. 2007 and Nyberg et al. 2013. Kouvonen et al. 2007 examined the relationship between job stress and smoking and alcohol consumption in public sector employees, while Nyberg et al. 2013 studied the association between job strain and cardiovascular disease risk factors. Focusing on Italy, Chiatti et al. 2010, analyzed exclusively the smoking habit, as Ficarra et al. 2011, but focalizing healthcare professionals.

Instead, the aims of this manuscript are: (1) to evaluate the relevance of considering several job-related variables when health topics are examined, (2) to explore how to analyze these job-related variables since an open-ended question is also present, (3) to discern sub-populations characterized by common SNAP and specific occupational types along with socio-economic variables. We want to give an initial overview of additional information related to work realities that policymakers should consider when targeting sub-populations for health prevention. The exploratory analysis proposed here was possible thanks to the availability of data coming from the cross-sectional Italian surveillance system PASSI (Baldissera et al., 2011). This system has been collecting data about lifestyle, behavioral risk factors, socio-demographic information, and self-diagnosed chronic diseases since 2007 with a high response rate, i.e., approximately 85%. Job-related information has always been collected with some further insights only in the last 10 years.

The analysis proposed is divided into several steps. First, text preprocessing is performed on the primary variable since it is an open-ended question where the interviewer asks about the respondent's job. Then, we perform a cluster analysis for mixed data based on medoids and Gower distances (Gower, 1971), considering the behavioral risk factors, socio-demographic variables, job sector, and classification as covariates. Finally, we analyze each cluster separately, understanding which type of job (coming from the first step) is most prevalent in the different clusters.

The paper is organized as follows. Section 2 describes the data analyzed and the related preprocessing steps. Section 3 briefly defines the clustering approach and related dissimilarity matrix. Finally, Section 4 is devoted to the results, while Section 5 is to the discussion.

2. PASSI Data

We analyze data from the Italian surveillance system PASSI, a sample survey that collects information about lifestyles, behavioral risk factors, socio-demographic information, and self-diagnosed non-communicable diseases. The population of reference is Italian adults ages 18 to 69. For additional information, please refer to Baldissera et al.2011 and the following webpage: https://www.epicentro.iss.it/passi/en/english. Our focus centers on the years preceding the COVID-19 pandemic, specifically from 2014 to 2019, during which job-related variables were recollected. Approximately 40% of the observations exhibit missing values concerning these job-related variables, whereas 88% refer to unemployed respondents. We then have a total of 129,100 observations (56% males). The variables related to socioeconomic variables (first two) and behavioral risk factors (last four) analyzed in the cluster analysis are defined in Table 1.

Table 1. Variables analyzed coming from the Italian surveillance system PASSI.

Variable	Description
Educational level	Low: if below high school; high: otherwise
Economic status	No: if the respondent easily meets financial needs; yes: otherwise.
Alcohol	No: never alcohol; yes: otherwise
Activity	Intense; moderate; no activity
Diet	No fruit; 1-2 portions; 3-4 portions; 5+ portions (per day)
Smoke	Smoker; ex-smoker; never smoke

Furthermore, we examine three variables related to the respondents' employment, defined below. The first stems from the query: "Can you tell me what you do for a living?". The second involves the classification of the declared job according to ISTAT (the Italian National Institute for Statistics) coding (https://professioni.istat.it/sistemainformativoprofessioni/cp2011/). Here, interviewers compile the information directly without querying respondents. As the response to the first job-related question is open-ended, an initial step involves proper text preprocessing to assess its coherence with the ISTAT job class declared by interviewers. Text preprocessing was conducted using the TextWiller R package (Solari et al., 2019), tailored for the Italian language. So, we perform web scraping on the ISTAT website (https://professioni.istat.it/ sistemainformativoprofessioni/ cp2011/) to correlate each job declared by the respondent with the corresponding ISTAT classification. We observed ≈ 60% coherence between the web scraping results and the class declared by the interviewer. After reviewing half of the mismatches, we consider the web scraping results more reliable and utilize them in the clustering step. Finally, the third job-related variable delineates nineteen occupational sectors: agriculture, industry (with specific subcategories, i.e., food, mechanical engineering, electrical and electronic, textile, chemical and ceramics, wood and paper, other), construction, energy-gas-

water-telecommunications, commerce and public establishments, transportation, banks and insurance, school-university, healthcare, public administration, business services, personal services, and law enforcement.

3. Cluster analysis

The joint analysis of several components related to SNAP risk factors at the same time presents challenges. Techniques like Principal Component Analysis (PCA) and factor analysis are commonly used to address high dimensionality. However, using PCA in a high-dimensional and heterogeneous variable set may hinder result interpretation. Factor analysis has limitations with categorical variables, such as the one analyzed in this manuscript. Instead, we opted for cluster analysis to group subjects with similar lifestyles and socioeconomic patterns.

We chose the k-medoids approach (Hastie et al., 2009) instead of other clustering methods, such as k-means and hierarchical clustering, since it works with any type of dissimilarity matrix and is robust with respect to the presence of outliers and noise. It is also less sensitive to initial cluster centers and the assumption of spherical cluster shapes.

As already pointed out, the variables under examination are qualitative ones. Therefore, the Gower distance (Gower, 1971) is utilized in the clustering approach. Let $j = 1, \dots, 9$ the index specifying the covariates defined in Section 2 plus the ISTAT and sector variables and $i, i' \in \{1, \dots, n\}$ the index denoting the respondents. Then, X_{ij} defines the value of the variable j at observation i. The Gower distance matrix $D \in \mathbb{R}^{n \times n}$ has elements $d_{ii'}$ defined as the average of single dissimilarities $d_{ii'j}$ (one for each variable). The single $d_{ii'j}$ is defined depending on the characteristic of the qualitative variable. We consider the educational level, economic status, alcohol variables as asymmetric binary variable while the physical activity, smoke and diet variables are defined as ordered categorical variables and the two job-related variables as categorical ones. Considering $d_{ii'j} = 1 - s_{ii'j}$, if the variable is asymmetric binary, we have $s_{ii'j}$ equals

$$s_{ii'j} = \begin{Bmatrix} 1 & if\ X_{ij} = X_{i'j} = 1 \\ 0 & otherwise \end{Bmatrix} \tag{1}$$

while if the variable is symmetric binary or nominal (i.e., more than 2 categories) we have

$$s_{ii'j} = \begin{Bmatrix} 1 & if\ X_{ij} = X_{i'j} \\ 0 & otherwise \end{Bmatrix}. \tag{2}$$

Finally, if the variables are ordered categorical variables, we have (Podani, 1999):

$$d_{ii'j} = \frac{|r_{ij} - r_{i'j}|}{\max(r_{ij}) - \min(r_{ij})}. \tag{3}$$

where r_{ij} denotes the rank of X_{ij}. Having the dissimilarity $D \in \mathbb{R}^{n \times n}$ the k-medoids clustering approach find $k \in \{1, \dots, n\}$ group of observations minimizing the sum of pairwise dissimilarities within the clusters.

4. Results

We imposed an a priori minimum number of three clusters to avoid reducing the complexity of the observations into a binary category. The silhouette index (Hastie et al., 2009) is used as an internal validation index to estimate the optimal number of clusters. Two clustering analyses were performed, one for the female population and one for the male population. The respondents were selected from among those between 29 and 50 years old.

4.1. Male population

The silhouette index equals 0.12, i.e., 5 is the optimal number of clusters. Figure 1 shows the bar plots for each variable defined in Table 1, while we comment below some of them in detail.

The cluster with the lowest risk of SNAP is the first one characterized by intellectual, scientific, and highly skilled professions (i.e., ISTAT code 2 and labor sector named "services to people"). Looking at the open-ended question, the five most frequent jobs are engineer, teacher, lawyer, office employee, and doctor. Clusters 2 and 4 are characterized by the same job ISTAT classification (laborers, artisans, and farmers) even if they present different scenarios in terms of SNAP and socio-economic variables. However, looking at the labor sector variable and the open-ended question, we found that the cluster with the worst situation is the construction sector, while cluster 4 is characterized by blue-collar and skilled workers in the engineering sector. This is an example of how the combination of the three job-related variables analyzed can give insight into the connection between work environment and health.

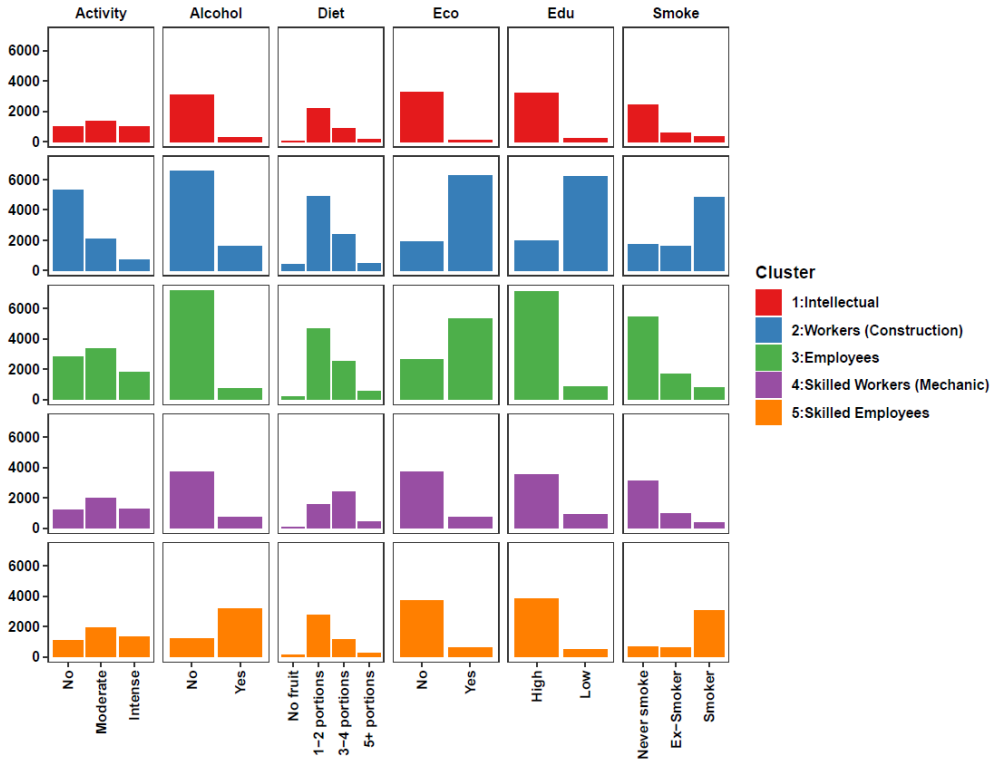


Figure 1. Frequency bar plots of variables defined in Table 1 in each cluster for the male population.

4.2. Female population

The optimal number of clusters is 8 (silhouette index equals 0.14). As in the previous subsection, we report in Figure 2 the bar plots for each variable defined in Table 1, while some clusters are examined in detail, analyzing the three job-related variables and the resulting medoids.

Clusters 1, 2, and 7 emerge as particularly vulnerable. The first cluster predominantly features roles within businesses (e.g., cashiers and shop assistants) and is associated with lower education levels, smoking, and limited physical activity. In contrast, the second cluster is centered around specialized healthcare professions, marked by both smoking and alcohol consumption. The seventh cluster is linked to public administration and predominantly comprises office workers, with absence of physical activity being the primary risk factor. Analyzing the female sub-population proves challenging because 75% of declared jobs are categorized as "employees." However, the other two job-related variables, ISTAT and sector, offer additional insights into the variability within this job type. For instance, clusters 5 and 7 feature the "office employee" job, with the former being associated with no economic problem

and the latter linked to the public administration sector and economic problem. This underscores again the significance of considering all three job-related variables in future studies.

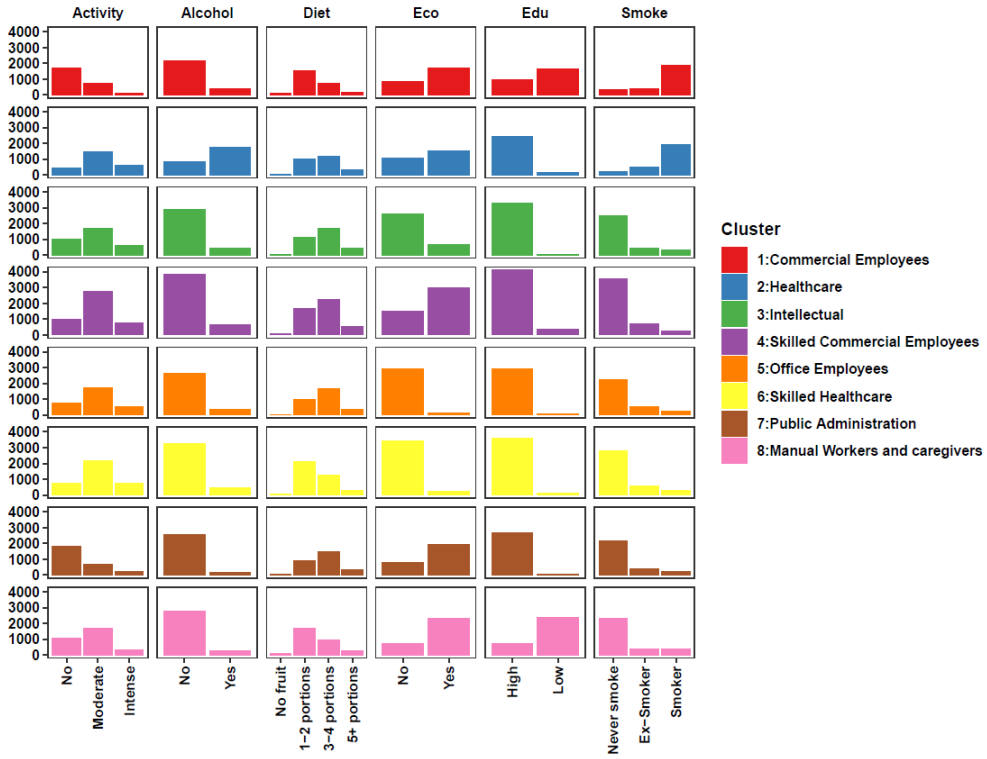


Figure 2. Frequency bar plots of variables defined in Table 1 in each cluster for the female population.

5. Discussion

This study explores the link between behavioral risk factors and socio-economic variables, focusing on detailed job types in Italy. The use of the Italian surveillance system PASSI data seems promising, offering valuable information (when adequately analyzed) that is difficult to gather in other ways. Here, we can analyze three job-related variables (i.e., the specific job coming from an open-ended question, ISTAT classification, and occupational sector), giving different insights into Italy's work realities. We perform a cluster analysis to identify sub-populations with distinct characteristics. Results reveal associations between certain occupations, such as building industry roles for males and healthcare professions for females, and higher levels of behavioral risk factors. The study emphasizes the importance of considering

the three job-related variables in understanding behavioral risk, providing valuable information for targeted health interventions based on specific populations. In particular, thanks to proper text preprocessing and web scraping, the open-ended question gives valuable detailed insights into the job realities of the sub-populations defined by the cluster analysis. However, some criticalities remain and need further investigation. In particular, in the female population, the "employee" job (coming from the open-ended question) is the most prevalent one as well as the most variable in terms of socio-economic work reality. Besides limitations and the need for further analyses, this approach fills the gap of information on health and work environments in Italy, offering this as an example also for other countries in which similar Risk Factors Surveillance Systems are running.

References

Baldissera S, Campostrini S, Binkin N, Minardi V, Minelli G, Ferrante G, Salmaso S. Features and initial assessment of the Italian behavioral risk factor surveillance system (PASSI), 2007–2008. Prev Chron Dis 2011;8(1).

Chiatti, C., Piat, S. C., Federico, B., Capelli, G., Di Stanislao, F., Di Giovanni, P., ... & Manzoli, L. (2010). Cigarette smoking in young-adult workers: a cross-sectional analysis from Abruzzo, Italy. Italian Journal of Public Health, 7(3).

Ficarra, M. G., Gualano, M. R., Capizzi, S., Siliquini, R., Liguori, G., Manzoli, L., ... & La Torre, G. (2011). Tobacco use prevalence, knowledge, and attitudes among Italian hospital healthcare professionals. *European journal of public health*, 21(1), 29-34.

Flaskerud, J. H., DeLilly, C. R., & Flaskerud, J. H. (2012). Social determinants of health status. *Issues in mental health nursing*, 33(7), 494-497.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 857-871.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: Springer.

Kouvonen, A., Kivimäki, M., Väänänen, A., Heponiemi, T., Elovainio, M., Ala-Mursula, L., ... & Vahtera, J. (2007). Job strain and adverse health behaviors: the Finnish Public Sector Study. *Journal of occupational and environmental medicine*, 49(1), 68-74.

Minardi, V., Campostrini, S., Carrozzi, G., Minelli, G., & Salmaso, S. (2011). Social determinants effects from the Italian risk factor surveillance system PASSI. *International journal of public health*, 56, 359-366.

Noble, N., Paul, C., Turon, H., & Oldmeadow, C. (2015). Which modifiable health risk behaviours are related? A systematic review of the clustering of Smoking, Nutrition, Alcohol and Physical activity ('SNAP') health risk factors. *Preventive medicine*, 81, 16-41.

Nyberg ST, Fransson EI, Heikkilä K, Alfredsson L, Casini A, et al. (2013) Job strain and cardiovascular disease risk factors: meta-analysis of individual-participant data from 47,000 men and women. PloS one, 8(6), e67323.

Prochaska, J. J., & Prochaska, J. O. (2011). A review of multiple health behavior change interventions for primary prevention. *American journal of lifestyle medicine*, 5(3), 208-221.

Solari, D., Sciandra, A., & Finos, L. (2019). TextWiller: Collection of functions for text mining, specially devoted to the Italian language. *Journal of Open Source Software*, 4(41), 1256-1257.

World Health Organization. (2014). Global status report on noncommunicable diseases 2014 (No. WHO/NMH/NVI/15.1). World Health Organization.

CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17800

Structuring and extracting sustainability information from corporate websites SMEs: A pilot test on textile firms

Francisco Javier Rodríguez-Ruiz[1], **Ana Garcia-Bernabeu**[2]

[1]Department of Textile and Paper Engineering, Universitat Politècnica de València, Spain, [2]Department of Economics and Social Sciences, Universitat Politècnica de València, Spain.

How to cite: Rodríguez-Ruiz, F. J.; Garcia-Bernabeu, A. 2024. Structuring and extracting sustainability information from corporate websites SMEs: A pilot test on textile firms. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17800

Abstract

In recent years, heightened awareness of environmental, social, and governance (ESG) issues has spurred a growing demand for sustainability-related data. While large corporations progress towards disclosing non-financial information, small and medium-sized enterprises (SMEs) face limitations due to the absence of standardized frameworks for reporting sustainable data. This paper aims to elucidate the process of developing a sustainability indicator framework by utilizing web-available information, encompassing the collection, processing, and analysis of ESG indicators. The structured extraction of ESG information has been assessed within a sample of textile SMEs in the Valencian Community over two years, aiming to provide an initial diagnosis of the quantity of sustainability reported information. The primary conclusion drawn is that companies are progressively incorporating such information into their web platforms, albeit without consistent coverage across all ESG dimensions and sub-dimensions.

Keywords: *Sustainability reporting; Web-based information; SMEs (Small and Medium-sized Enterprises); Sustainability indicator framework; Textile*

1. Introduction

In the past few years, there has been a significant increase in awareness regarding environmental, social and governance (ESG) concerns, leading to a substantial surge in the demand for sustainability data. Unlike larger corporations, Small and Medium Enterprises (SMEs) may not have standardized reporting frameworks for sustainability disclosure. The absence of universal metrics and reporting standards makes it challenging to compare and benchmark sustainability performance across different SMEs (Pranugrahaning et al., 2021, Martins et al., 2022).

In the lack of a standardized framework for assessing corporate sustainability, many companies, including SMEs, have chosen to report their engagement in sustainable practices through their corporate website (Cruciata et al. 2023; Palma et al. 2022; Wanderley et al., 2008; Lodhia, 2010). This digital footprint, when monitored for sustainability performance, becomes a powerful tool for gaining insights into company behaviors and practices. With recent advancements in Natural Language Processing (NLP) technologies, the ability to extract meaningful information from this digital trail has been greatly improved (Blazquez and Domènech, 2008, Luccioni et al., 2020).

The aim of this paper is to propose a framework of sustainability indicators for SMEs that permits the extraction of web-based information on their environmental, social and good governance practices. The selected indicators have been obtained by adapting the proposal of the Bank of Spain's BELAb project for large companies (Fernández-Rosillo San Isidro et al., 2023), as well as other reference reports focused on the case of SMEs. Once the indicators have been defined, an initial diagnosis will be made of the information disclosed through the web in a sample of textile companies in two periods, 2021 and 2024, to see how the disclosed sustainability information via their websites has evolved. This proposal is an innovative approach to adapt a sustainability indicator framework originally designed for large companies to SMEs. The use of NLP to extract and analyze web-based sustainability information is a creative and practical solution to the problem of non-standardized information in SMEs.

The paper is organized as follows. Section 2 presents the proposed monitoring framework by exploring and identifying relevant sustainability indicators for SMEs. Section 3 explains the methodology used to extract the information about the selected indicators. Next, in Section 4, we present an application of extracting and structuring ESG information in a sample of SME textiles companies in Spain. Finally, the paper ends with conclusions and future line of research.

2. Exploring and determining the relevant indicators for SMEs

Regulation of non-financial reporting differs by country and jurisdiction, but there has been a growing interest around the world in promoting non-financial disclosure. In the European Union, for example, Directive 2014/95/EU on non-financial disclosure (European Union, 2014) and diversity for large companies was implemented, requiring certain companies to report on environmental, social and personnel, human rights, and anti-corruption issues.

Although progress has been made in the disclosure of non-financial information currently, SMEs are not obligated to submit a Non-Financial Information Statement (NFIS). Nevertheless, regulatory frameworks are undergoing changes, and SMEs may be mandated to do so in the future. Our analysis encompassed the list of preliminary indicators proposed by Fernández-Rosillo San Isidro et al., (2023) which has been complemented taking into account several international ESG standards, with a particular focus on delving into the technical documentation

of the Global Reporting Initiative (GRI). This focus on GRI was driven by the significant number of SMEs companies that choose to report according to this standard. Table 1 presents an initial compilation of 37 ESG indicators distributed across three main ESG dimensions, each further classified into ten subtypes. The environmental dimension encompasses distinct groups such as Energy, Water, Greenhouse Gases, Waste, and Environmental Policies. Within the social dimension, indicators are structured into three groups: Employees, Diversity, and Society. Lastly, the governance dimension is characterized by a set of specific indicators related to corporate governance and corruption and bribery.

3. Materials and methods

3.1. Methodology

The process followed for information extraction is divided into three stages.

First stage: Definition of keywords associated to each indicator. Following the proposal suggested in Section 2, a keyword dictionary derived from the stem of words for each ESG indicator is built to track the presence of the text on the company's website.

Second stage: Data extraction process with a procedure like the one described in Blázquez et al. (2018) and Crosato et al. (2021). To analyze the disclosure of information about an indicator this stage is designed to answer the question: "Is this text about the label X in the ESG indicator included in the web? The answer to this question is an indicator of the company's awareness about the ESG indicator. Next, the number of instances in which a label appears in association with an indicator in the companies' web pages is transformed into values of "1" if it appears, or "zero" otherwise.

Third stage: A Wilcoxon signed-rank test was employed to evaluate differences between the years 2021 and 2024 in the median frequency of keywords appearing on company websites. This non-parametric approach was chosen because it makes no assumptions about the underlying data distribution. Null Hypothesis (H_0): There is no statistically significant difference in the median frequency of keyword occurrence between the years 2021 and 2024; Alternative Hypothesis (H_1): There is a statistically significant difference in the median frequency of keyword occurrence between the years 2021 and 2024.

Table 1. Selection of ESG indicators for sustainability reporting in SMEs. ***Note.*** Developed based on Corral-Lage et al. (2021) & Fernández-Rosillo San Isidro et al. (2023).

Type	Subtype	Identifier	Indicator
E	Energy	E01_ENCO	Energy consumption
		E02_ROEC	Reduction of energy consumption
		E03_RETE	Percentage of renewable energy relative to total energy consumed
	Water	E04_WACO	Water consumption
	Green House Gases	E05_GHEI	GHG emissions intensity
		E06_GHER	GHG emissions reduction
	Waste	E07_WAGE	Waste generated
		E08_HAWW	Hazardous waste
		E09_MAWA	Managed waste
		E10_REWA	Reused waste
	Environmental Policies	E11_CIEC	Circular economy
		E12_ENPO	Environmental policy
		E13_ISCO	Regulatory compliance
S	Employees	S01_EMTR	Employee training
		S02_DISA	Disability
		S03_JOST	Job stability
		S04_EMTU	Employee turnover
		S05_NUOD	Number of dismissals
		S06_WLBP	Work-life balance policies
		S07_OCRI	Occupational risk
		S08_JOTE	Job seniority
		S09_EMPL	Employees
	Diversity & equality	S10_GEDI	Gender diversity
		S11_AGPG	Average gender pay gap
		S12_EQPL	Equality plan
		S13_DIPL	Diversity plan
	Society	S14_HASP	Health and safety policy
		S15_HURP	Human rights policy
		S16_SUPA	Supplier payments
		S17_SUCH	Supply chain
G	Corporate Governance	G01_AVBR	Average board remuneration
		G02_BOME	Board Meetings
		G03_GDIT	Gender Diversity in the Board
	Corruption and bribery	G04_COAB	Corruption and bribery
		G05_CRPP	Crime prevention policy
		G06_NCAB	Number of corruption and bribery reports
		G07_WHCH	Whistleblower channel

Table 2. Example of labels associates to subtype dimension "Energy".

Type	Subtype	Dictionary of labels (Spanish)
E	Energy	Consumo de energía; GRI 302; Consumo energético; Economía Circular; Uso de agua; Huella hídrica, Residuos; Reciclaje; Energía Renovable, …

3.2. Data

The sample for this study covers 215 textile SMEs located in the Comunidad Valenciana region in Spain, with data for the years 2021 and 2024 considered. The sample of official websites of companies was retrieved from the SABI (Iberian Balance Sheet Analysis System) database after a selection and filtering process of active companies in the Comunidad Valenciana textile sector in both years and with a single URL. As for the sustainability variables they were extracted from the websites of the companies after crawling the complete website in both years.

4. Results

The Wilcoxon test revealed that the p value obtained (0.001) is less than the significance level ($\alpha = 0.05$), leading to the rejection of the null hypothesis. Therefore, it is concluded that there is statistical evidence to affirm that the medians of the frequency of appearance of the keyword between the years 2021 and 2024 are different. This result suggests a significant change in the disclosure of sustainability information on company websites during that period, indicating an increase in concern about sustainability.

Looking at the disclosed information on sustainability in our sample for both years in Table 3, the first conclusion to be drawn is that there has been a 27,72% increase in information in 2024 compared to 2021.

Table 3. Summary of ESG information retrieved by type.

Type	2021	2024	Increase
E	217	283	30,4%
S	306	366	19,6%
G	36	65	80,6%
Total	559	714	27,7%

Figure 1 shows the distribution of disclosed information by type (environmental, social or governance). Notice that, while there is an increase in the percentage of disclosed information

about environmental and governance dimensions, there is a small decrease in the percentage of reporting of social dimension.

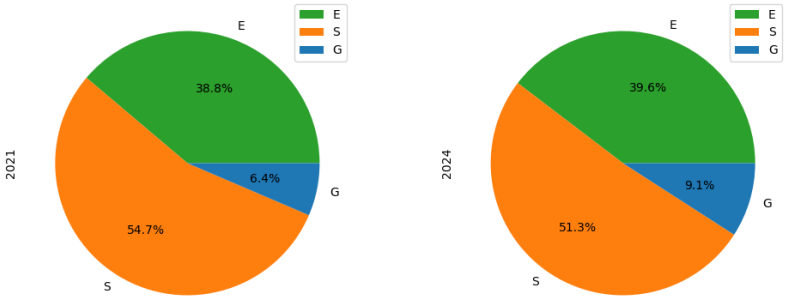


Figure 1. Benchmarking of information presented on the web by type (ESG).

Regarding the subtype of information included in the group of environmental indicators (see Figure 2), there is a greater presence of information contained in the subcategories of energy policies, energy and water. On the other hand, there is almost no information on gas emissions and water consumption. As for the set of social indicators, information on employee welfare, equality and diversity issues is predominant, while website disclosed information on the company's impact on society is less present. Finally, in the group of governance indicators, there is a growing interest in disclosure information related to whistleblowing and corruption channel.

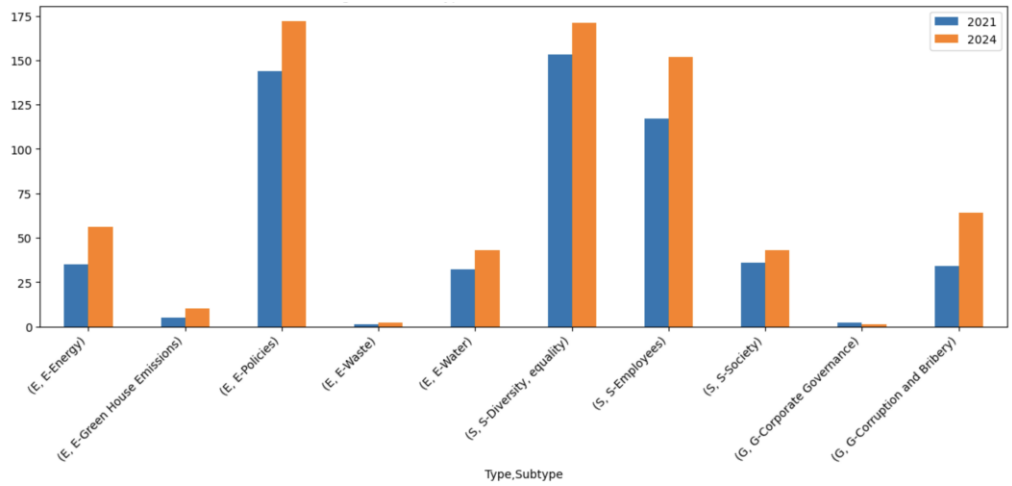


Figure 2. Benchmarking of ESG information presented by subtype.

5. Conclusions

In this paper we have developed a framework for analyzing complementary sustainability information for SMEs based on unconventional data extracted from company's websites. The proposed framework, adapted from the Bank of Spain's BELAb project and other relevant reports, serves as a valuable tool for analyzing sustainability indicators tailored to the unique characteristics of SMEs. The research successfully leverages Natural Language Processing (NLP) technologies to extract meaningful sustainability information from the digital footprint left by companies on their websites.

The distribution of disclosed information by subtype provides a more detailed overview, emphasizing the importance of certain indicators within each dimension. For instance, the paper highlights the prominence of data related to energy policies, energy, and water within environmental indicators, as well as the growing interest in governance indicators related to whistleblowing and corruption channels.

The findings underscore the progress made by textile SMEs in incorporating sustainability information into their digital presence, while also pointing towards areas where improvements and standardization are necessary for a more transparent and comprehensive disclosure landscape. Future research should build upon these insights. Furthermore, this exploration is encouraged to encompass a more extensive sample of companies in other regions and industries, spanning multiple years, to provide a comprehensive understanding of evolving practices. In addition, future research could enhance the process of keyword extraction for ESG indicators by using machine learning techniques.

References

Blazquez, D. and Domenech, J. (2018). Big data sources and methods for social and economic analyses. *Technological Forecasting and Social Change, 130*, 99–113.

Corral Lage, J., García Delgado, S., Ipiñazar Petralanda, I., Peña Miguel, N., Saitua Iribar, A. 2021. Guía para la emisión y verificación de información sostenible a través de indicadores medioambientales, sociales y de gobernanza para PYMEs. BNFIX GLOBAL, S.L. pp. 66. Retrieved March 03, 2024, from https://www.bnfix.com/wp-content/uploads/2021/11/GUIA_BNFIX_impreso-1.pdf

Crosato, L., J. Domènech, and C. Liberati (2021). Predicting SME's default: Are their websites informative? *Economics Letters*, 204:109888.

Cruciata, P., Pulizzotto, D., Héroux-Vaillancourt, M., & Beaudry, C. (2023). 0-shot text classification for web-based environmental indicators: Pilot study on B-Corp data. *5th International Conference on Advanced Research Methods and Analytics (CARMA2023)*. http://dx.doi.org/10.4995/CARMA2023.2023.16463

European Union. (2014). Directive as regards disclosure of non-financial and diversity information by certain large undertakings and groups, 2014/95/EU. Retrieved March 03, 2024, from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32014L0095

Fernández-Rosillo San Isidro, B., Koblents Lapteva, E., & Morales Fernández, A. (2023). Micro-database for sustainability (ESG) indicators developed at the Banco de España (2022). *Notas estadísticas/Banco de España*, *17*.

Lodhia, S. K. (2010). Research methods for analysing world wide web sustainability communication. *Social and Environmental Accountability Journal*, *30*(1), 26–36.

Luccioni, A., Baylor, E., & Duchene, N. (2020). Analyzing sustainability reports using natural language processing. *arXiv preprint arXiv:2011.08073*.

Martins, A., Branco, M. C., Melo, P. N., & Machado, C. (2022). Sustainability in small and medium-sized enterprises: A systematic literature review and future research agenda. *Sustainability*, 14(11), 6493.

Palma, M., Lourenço, I. C., & Branco, M. C. (2022, October). Web-based sustainability reporting by family companies: the role of the richest European families. In *Accounting Forum* (Vol. 46, No. 4, pp. 344-368). Routledge.

Pranugrahaning, A., Donovan, J. D., Topple, C., and Masli, E. K. (2021). Corporate sustainability assessments: A systematic literature review and conceptual framework. *Journal of Cleaner Production, 295, 126385.*

Wanderley, L. S. O., Lucian, R., Farache, F., and de Sousa Filho, J. M. (2008). CSR information disclosure on the web: a context-based approach analysing the influence of country of origin and industry sector. *Journal of business ethics*, 369–378.

Topic Modelling with Constructivist Grounded Theory: A Way of Big Textual Data Analysis for Theory Building

Eyyub Can Odacioglu[1], Lihong Zhang[1], Richard Allmendinger[2], Azar Shahgholian[3]

[1]Department of Engineering Management, the University of Manchester, UK, [2]Alliance Manchester Business School, the University of Manchester, UK, [3]Liverpool Business School, Liverpool John Moores University, UK.

How to Cite: Odacioglu, E.C.; Zhang, L.; Allmendinger, R.; Shahgholian, A. 2024. Topic Modelling with Constructivist Grounded Theory: The Way of Big Textual Data Analysis for Theory Building In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. p. 407. https://doi.org/10.4995/CARMA2024.2024.19017

Abstract

There is a growing demand for methodological plurality, especially with the emergence of Machine Learning (ML) techniques for analysing big textual data. To address this need, our study introduces a novel methodology that combines a ML technique with a traditional qualitative approach to reconstruct knowledge from existing publications. With its pragmatist and abductive stance, it allows for human-machine interaction. The method employs Topic Modelling (TM), an ML technique, to facilitate Constructivist Grounded Theory (CGT). A four-step coding process (Raw Coding, Expert Coding, Focused Coding, and Theoretical Coding) is implemented to ensure procedural and interpretive rigor. To present this approach, we collected data from an open-source professional project management community website and illustrated the research design, data collection, and data analysis processes leading to theory development. The results revealed the potential of this novel methodology to extract latent meanings and reveal phenomena within published data, thereby offering a new avenue for academics to develop potential theories in various fields.

Keywords: *Big Data; Grounded Theory; Machine Learning; Substantive Theory Building; Topic Modelling*

Boosting XGBoost and Neural Networks - Using the Panel Dimension to Improve Machine-Learning-Based Forecasts in Macroeconomics

Jonas Dovern[1], Johannes Frank[2]

[1]School of Business, Economics and Society, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany, [2]School of Business, Economics and Society, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany.

How to cite: Dovern, J.; Frank, J. 2024. Boosting XGBoost and Neural Networks – Using the Panel Dimension to Improve Machine-Learning-Based Forecasts in Macroeconomics. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. p. 408. https://doi.org/10.4995/CARMA2024.2024.19017

Abstract

The short time dimension of commonly used macroeconomic data sets presents challenges for the estimation of machine learning models designed for real-time business cycle monitoring. In this paper, we consider panel data to increase the data set available for training and nowcasting US unemployment using extreme gradient boosting and neural networks. The underlying idea is that dynamics between variables and across time at the state level are similar to each other and to the dynamics at the national level. We use data pooling in combination with weight sharing that accommodates some cross-sectional heterogeneity. This approach facilitates parameter regularization and safeguards against overfitting. We find that this "soft" pooling approach improves forecast accuracy at the national level and reduces both the variance and the mean of the RMSE distribution across states. Thus, leveraging regional information in a panel data framework with suitable regularization techniques addresses data scarcity in macroeconomic nowcasting effectively.

Keywords: *nowcasting; unemployment; pooling; panel data; XGBoost; neural networks.*

Analysis of the trend of tourist visits through photographs uploaded on social media

María del Rocío Martínez-Torres[1], Myriam González-Limón[2], Francisco Javier Quirós-Tomás[1], Lourdes Cauzo-Bottala[1]

[1]Department of Management and Business Administration at Business Administration and Marketing, University of Seville, Spain, [2]Department of Economic Analysis and Political Economy, University of Seville, Spain.

How to cite: del Rocío Martínez-Torres, M.; González-Limón, M.; Quirós-Tomás, F. J.; Cauzo-Bottala, L. 2024. Analysis of the trend of tourist visits through photographs uploaded on social media. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. p. 409. https://doi.org/10.4995/CARMA2024.2024.19017

Abstract

The overall objective of this study is to investigate if it is possible to analyse the trend of tourist visits through photographs uploaded on social media. Three analysis have been developed using a database on two time series created from photographs uploaded to Flickr website from January to end of May in both 2019 and 2020: a descriptive analysis of the number of photographs, Pettitt's test and univariate boxplot or box-and-whisker plot. Our study confirms first, that the changes in the trend of photograph uploads to Flickr imply changes in the number of tourist visits, and second, he usefulness of analysing photography social media as a source of information to assess the effects of different events, expected and unexpected, on tourism.

Keywords: *Flickr; Pettitt's test; Tourist visits trends; geo-tagged photographs; World Heritage Sites*

#SDG5 – Social Media Intelligence analysis of Gender Equality

Enara Zarrabeitia-Bilbao, Izaskun Álvarez-Meaza, Maite Jaca-Madariaga, Rosa María Rio-Belver

Industrial Organization and Management Engineering Department, University of the Basque Country, Spain.

How to cite: Zarrabeitia-Bilbao, E.; Álvarez-Meaza R.; Jaca-Madariaga M; Rio-Belver RM. 2024. #SDG5 – Social Media Intelligence analysis of Gender Equality. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. p. 410. https://doi.org/10.4995/CARMA2024.2024.19017

Abstract

This study focuses on the digital conversation related to SDG 5 (Gender Equality), between 2016 and 2022 on Twitter, analysing 394,494 tweets. After obtaining, cleaning and refining data, the Social Media Intelligence analysis of Gender Equality was carried out through Social Network Analysis (SNA) methods and Artificial Neural Networks (ANN) models. The results obtained revel that the conversation is intermittent and phenomena such as International Women's Day act as catalysts for digital conversation. It is also worth highlighting the presence of anomalous behaviour related to certain profiles that could be attributable to bots. In this respect, future studies should corroborate this fact and be able to detect certain patterns of action for this type of profiles. Finally, during the entire period analysed, the conversation revolves around positive terms, which denotes a clearly constructive, nonpolemic tone.

Keywords: *Gender Equality; SDG 5; Social Media Intelligence; Social Network Analysis; Artificial Neural Networks; Sentiment Analysis*

Do websites provide information about innovation activities?

Agapito Emanuele Santangelo
Department of Economics, University of Molise, Italy.

How to cite: Santangelo A. E. 2024. Do websites provide information about innovation activities?. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. p. 411. https://doi.org/10.4995/CARMA2024.2024.19017

Abstract

How can we measure the most recent technological developments in order to determine their economic implications? Throughout the initial phases of their life cycles there is only limited data available to track technologies (Raj and Seamans 2018). The purpose of this work is to present the use of on-line indicators derived from corporate websites as a means of assessing corporate activities for firms granted with subsideis related to innovative activities. Leveraging online indicators obtained through web scraping this study aims to enhance the understanding of companies behaviors. The research include both offline data from the balance sheets of companies and online data obtained through web scraping of company websites, as the cornerstone of the analysis. Also, the Wayback Machine, hosted by the Internet Archive, serves as a useful tool for monitoring websites over time offering a vast archive of more than 26 years of web history.

Keywords: *Web Scraping; Subsidies; Innovation.*

Male Supremacy Online An Investigation of Incel Ideology Through Qualitative Content Analysis and Active Machine Learning

Mara Weber
SOCIUM, University of Bremen, Germany.

How to cite: Weber, M. 2024. Male Supremacy Online. An Investigation of Incel Ideology Through Qualitative Content Analysis and Active Machine Learning. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. p. 412.
https://doi.org/10.4995/CARMA2024.2024.19017

Abstract

This study aims to comprehensively examine the misogynistic incel ideology with a mixed-method framework based on 4.7 Million posts from the online plattform incels.is. The online subculture reinforces male supremacist beliefs through their black-pill ideology, which is a biological deterministic idea attributing mate choice solely to biological traits. Misogynistic incels pose threats to the public safety and gender equality. The study thoroughly examines the ideological dimensions and their interconnectedness. Integrating qualitative content analysis with active machine learning achieves a more holistic understanding of misogynistic ideology than an exclusive machine learning approach.

Methodically, the study argues in favor of mixed-method approaches between qualitative social research and machine learning to address informational gaps, such as interpretational ambiguities and arbitrary analytical choices inherent in machine learning models. This approach provides more robust measurements compared to those using machine learning alone.

Keywords: *Computational social science; Misogyny; Male supremacy;Mmixed-methods; Text-as-data.*