UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ARMA 2020

July 8-9, 2020 Valencia, Spain

3rd International Conference on Advanced Research Methods and Analytics

ADE FACULTAD DE ADMINISTRACIÓN Y DIRECCIÓN DE EMPRESAS. UPV

Departamento de Economía y Ciencias Sociales

devstat

Congress UPV

3rd International Conference on Advanced Research Methods and Analytics (CARMA 2020)

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

Publisher

2020, Editorial Universitat Politècnica de València
www.lalibreria.upv.es / Ref.: 6563_01_01_01

Cover design by Gaia Leandri

ISBN: 978-84-9048-832-4 (print version)
Print on-demand

DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11920

Preface

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

Abstract

Research methods in economics and social sciences are evolving with the increasing availability of Internet and Big Data sources of information. As these sources, methods, and applications become more interdisciplinary, the 3rd International Conference on Advanced Research Methods and Analytics (CARMA) is an excellent forum for researchers and practitioners to exchange ideas and advances on how emerging research methods and sources are applied to different fields of social sciences as well as to discuss current and future challenges. This edition was celebrated virtually because of the COVID-19 outbreak.

Keywords: *Big Data sources, Web scraping Social media mining, Official Statistics, Internet Econometrics, Digital transformation, global society.*

1. Preface to CARMA2020

This volume contains the selected papers of the Third International Conference on Advanced Research Methods and Analytics (CARMA 2020) virtually hosted by the Universitat Politècnica de València, Spain during 8 and 9 July 2020. Despite the COVID-19 outbreak, This third edition consolidates CARMA as a unique forum where Economics and Social Sciences research meets Internet and Big Data. CARMA provides researchers and practitioners with an ideal environment to exchange ideas and advances on how Internet and Big Data sources and methods contribute to overcome challenges in Economics and Social Sciences, as well as on the changes in the society because of the digital transformation.

The selection of the scientific program was directed by Maria Rosalia Vicente, who led an international team of 67 scientific committee members representing institutions worldwide. Following the call for papers, the conference received 94 paper submissions from all around the globe. All submissions were reviewed by the scientific committee members under a double blind review process. Finally, 47 papers were accepted for oral presentation during the conference. This represents an overall paper acceptance rate of 50%, ensuring a high quality scientific program. It covers a wide range of research topics in Internet and Big Data, including official statistics, web scraping, search engine data, industry adoption, sentiment analysis, geospatial data or consumer behavior, among others. Additionally, 12 posters with promising work-in-progress research were selected for presentation during the conference.

Apart from the regular scientific sessions, the keynote speech was contributed by Pablo de Pedraza, who talked about “The semicircular flow of the data economy” and provided a unique view from his position in the Monitoring, Indicators and Impact Evaluation Unit at the Joint Research Centre of the European Commission.

The conference organizing committee would like to thank all who made this third edition of CARMA a great success. Specifically, thanks are indebted to the authors, scientific committee members, reviewers, invited speaker, session chairs, presenters, sponsors, supporters and all the attendees. Our final words of gratitude must go to the Faculty of Business Administration and Management of the Universitat Politècnica de València for supporting CARMA 2020.

2. Organizing Committee

General chair

Josep Domènech, Universitat Politècnica de València

Scientific committee chair

María Rosalía Vicente, Universidad de Oviedo

Local organization

Eduardo Cebrián
Mónica Costa Alcaina
Eduardo Torán

3. Sponsors and Supporters

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

4. Scientific committee

Anto Aasa, University of Tartu
Fernando Almeida, University of Porto & INESC TEC
Helena Alves, University of Beira Interior
María del Pilar Ángeles, Universidad Nacional Autónoma de México
Concha Artola, Banco de España
Nikolaos Askitas, IZA – Institute of Labor Economics
Seyhmus Baloglu, University of Nevada
Catherine Beaudry, Polytechnique Montreal
Silvia Biffignandi, University of Bergamo
Federico Botta, University of Exeter
Petter Bae Brandtzaeg, University of Oslo/SINTEF digital
Levent Bulut, Valdosta State University
Ludovic Calès, European Commission, JRC
José Luis Cervera, DevStat
Cihan Cobanoglu, University of South Florida
Marisol B. Correia, ESGHT-Algarve University & CiTUR
Piet J.H. Daas, Statistics Netherlands/Eindhoven University of Technology
Stefano De Marco, University of Salamanca

Rosa Rio-Belver, Universidad del Pais Vasco
Anna Rosso, University of Milano DEMM
Pål Sundsøy, NBIM
Sergio Toral Marin, Universidad de Sevilla
Konstantinos P. Tsagarakis, Democritus University of Thrace
Joonas Tuhkuri, MIT
Tiziana Tuoto, Istat Italian National Institute for Statistics
Antonino Virgillito, Italian Revenue Agency
Maro Vlachopoulou, University of Macedonia/Greece
Martin R. Wolf, University of Applied Sciences Aachen
Selim Zaim, Istanbul Sehir University

Index

Full papers

Abstracts

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11315

A method for determining the emergence level of transformer technologies for green energy applications

Gaizka Garechana, Rosa Río-Belver, Enara Zarrabeitia, Izaskun Álvarez-Meaza
Department of business management, University of the Basque Country, Spain.

Abstract

Solid State Transformers (SST) are the result of merging the power electronics possibilities for voltage and frequency control with high-frequency transformers, and are expected to be a key component for enabling some important features that future energy grids must possess: reversibility, stability, modularity and compactness, among others. In addition to this, the possibilities of SSTs can be enhanced with advanced semiconductor materials such as Silicium Carbide (SiC), considerably improving the voltage and frequency ranges of this devices. This study aims at developing a quantitative method for characterizing the emergence level of SSTs and SiC-based transformers in three areas where these technologies can have a sizable impact: photovoltaic (PH) and eolic (EO) energy production and electric vehicle (EV) appliances. Results show that PH area will probably outpace the EO area in both technologies, but the attention of the scientific community may be shifting from PH in the SiC-based transformer technology. EV applications are, on average, closer to the life cycle's exponential growth stage than PH and EO areas, so it seems reasonable to expect a comparatively faster increase of both scientific and technology development activity in this field.

Keywords: *SST; Silicium Carbide; Technology Forecasting; Emergent technologies.*

1. Introduction

Electrical transformers are devices that allow changing the voltage of electric current, a well-known application of transformers is that of increasing the voltage of the alternating current (AC) generated in power stations from low to high (step-up transformer), in order to increase the voltage and reduce the current, given that less current means that less energy is lost when transporting said current to the customers. High voltage current, however, is dangerous for typical household purposes, so voltage must be reduced back to safe levels (step-down transformer) before its final use. The conventional transformer presents some drawbacks for its deployment in local, decentralized, renewable energy grids. First, conventional transformers are too big for many of these applications. Second, transformers are one-way tools, suited for energy distribution systems designed around big, centralized power plants, which comes in stark contrast with the operational aspects of the smart and clean technologies that are expected to substantially increase their share in the energy mix of the future (Roberts, 2018). There is a need for small and flexible transformation systems that can also deal with energy storage systems that work on direct current (DC).

Solid State Transformers (SST) are the result of merging the power electronics possibilities for voltage and frequency control with high-frequency transformers. One of the core tasks of power electronics in these devices is to increase the typical current frequency coming from the grid (50 Hz in Europe) to a range between 10 and 20 KHz in order to feed a high-frequency transformer that could be 20% smaller than a conventional transformer. This can be achieved using conventional silicon-based insulated-gate bipolar transistors (IGBT), at the expense of the reliability and limited handling of voltage (6.5 Kv). New semiconductor materials such as the silicium carbide (SiC) address these shortcomings, enabling SSTs to work at higher voltages and very high frequencies, thus achieving the maximum reduction in transformer size (Bhattacharya, 2017). The flexibility brought by the power electronics also allows the adjustment of the SST to frequent shifts in voltage and the requirements of a smart energy-management system (Abu-Siada, Budiri, & Abdou, 2018).

The three-module approach proposed by Bhattacharya (2017) offers direct DC connection in the components of the SST, allowing to build direct interfaces with solar or other renewable energy technologies, thus avoiding extra DC – AC conversion steps and consequently, improving the efficiency. The possibilities of this multi-port structure go even further: Three module SiC-based SST systems can also be used to provide high voltage DC connection ports for electric vehicle (EV) quick chargers, using compact devices. The reversibility of the system could even allow using the local fleet of electric cars as a storage system for backing up the grid when necessary (Ronanki, Kelkar, & Williamson, 2019).

2. Research goals

The goal of the research presented in this paper is to develop a method suitable for determining the relative (since all the technologies studied in this paper are considered to be "emergent" according to the industry consensus) maturity of the applications of SST technology in the areas of photovoltaic energy (PH), eolic energy (EO) and the electric vehicle (EV), where the new transformer technologies are deemed to have a high impact.

The above mentioned method will also be applied to transformer technologies based on the advanced semiconductor material SiC, in order to analyze the penetration of this semiconductor in the transformer industry, in the PH, EO and EV areas.

3. Methodology

The basic premise underlying our methodology is that the emergence stage of the technology life cycle corresponds with the exponential growth stage of the logistic growth curve (Kucharavy & De Guio, 2015) and consequently, the current degree of development and future prospectives of an emergent technology can be characterized by fitting the data corresponding to that technology to an exponential model.

3.1. Data retrieval and subsetting

The present study uses scientific publication and patent data for the characterization of the developments taking place in a technological field. According to the linear model of innovation (Godin, 2006), advances in scientific activity should come before the development efforts (patents) at the organizations. Data was retrieved by running the following queries on Scopus database (scientific publications):

- SST technology: SST: TITLE-ABS-KEY (solid W/0 state W/0 transformer)
- SiC in transformer technology: (TITLE-ABS-KEY ("SILICON CARBIDE" OR "SILICIUM CARBIDE" OR carborundum OR sic) AND TITLE-ABS-KEY (transformer))

…and their approximate equivalents in Patseer (patents) database:

- SST technology: (TACD:(SOLID wd0 STATE WD0 TRANSFORMER))
- SiC in transformer technology: ((TACD:(("silicon carbide" OR "silicium carbide" OR "carborundum") AND TRANSFORMER) AND AC:(H02M*)))

In order to accomplish the goals stated in section 0, data was subsetted by applying text mining techniques to the "description" field of patents and "title, abstract & keywords" fields of scientific publications. We assume that the presence of terms unequivocally associated with PH, EO or EV applications in these fields is an evidence that can be used for subsetting

our data in three subsamples corresponding to the aforementioned areas. This approach provides a total amount of twelve subsamples to perform our study: six subsamples corresponding to scientific activity (PH, EO, EV – Scopus) in both SST and SiC transformer technologies, and their equivalent in patenting activity, (PH, EO, EV – Patseer) in both SST and SiC transformer technology.

These being emergent technologies, we obtained very few data prior to year 2000, so the time interval was defined from 2003 to 2017 for SiC transformer data, and from 2011/12 to 2018 for SST data. We pragmatically selected the start of the interval by selecting the first year in which at least one record was obtained for two out of the three areas under analysis (PH, EO, EV), for both SST and SiC transformer technologies.

3.2. Fitting the data to the exponential curve

The number of publications and patents corresponding to the twelve subsamples is fitted to an exponential model, according to the following method:

First, we parametrize the equation $y = ax^b$ by taking logs $\log(y) = \log(a) + b * \log(x)$ and fitting a linear regression model to the log-transformed data . This will provide an initial estimation of a and b parameters that will be subsequently recalculated by using the nonlinear weighted least-squares (NLS) method on the data, as explained in Hastie (2017), using R. The total amount of patents and publications corresponding to each area (PH, EO, EV) will inform abut the diffusion achieved by the technologies under study (SST and SiC based transformers) while the parameter b will inform about the relative maturity of said technologies in each area, considering that the higher the b, the higher are the expectations about the success of the technology in a particular area, and the closer is that technology to the exponential growth phase of the technology life cycle.

4. Results

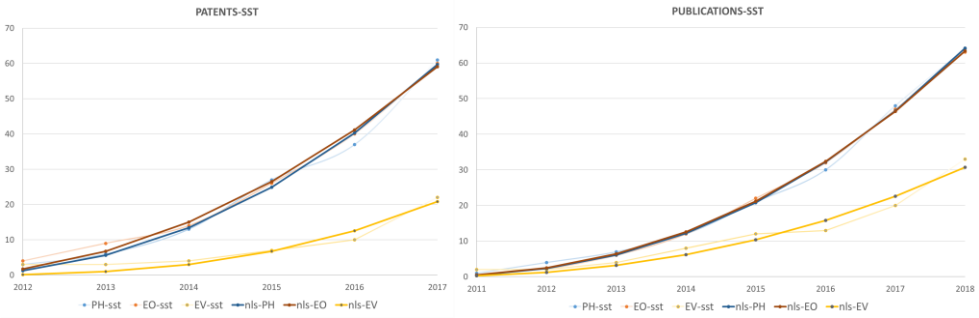

Figure 1 shows the accumulated patent and publication data corresponding to SST technology, as well as the results of fitting the model to the data:

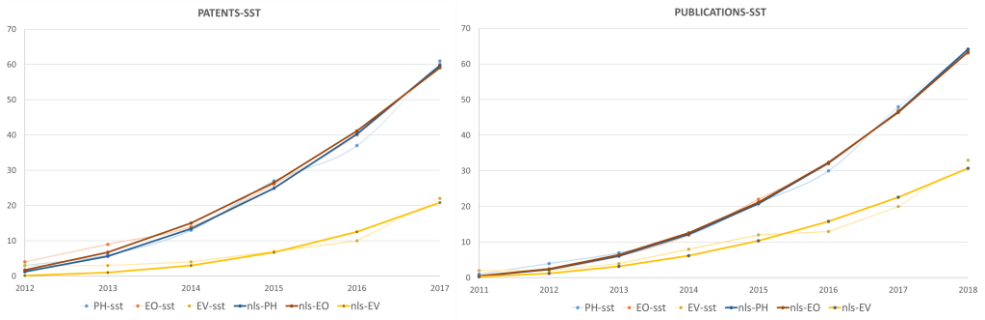

Figure 1. Data corresponding to SST technology (accumulated publications and patents). Colors indicate the application area (PH, EO, EV) and the thickness of the line indicates whether it shows raw data (thin) or the result of fitting the NLS model (thick).

The first thing we notice is that both data sources (publications and patents) show approximately the same starting point for the data, according to the criteria we exposed on section 0, and the patterns shown by data are also similar for both sources. The applications of SST to the EV area are fewer in number, and start to show a growth pattern later, when compared with PH and EO data. However, the *b* parameter is significantly higher for patent data in EV area, as can be seen in the results presented in Table 1.

Figure 2 shows the results corresponding to the SiC transformer technology:

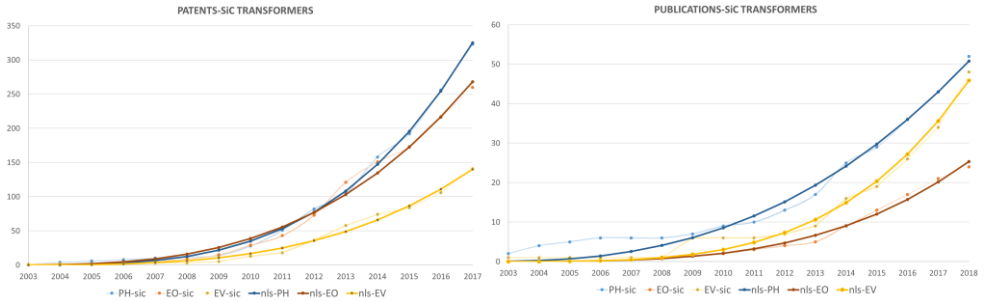

Figure 2. Data corresponding to SiC transformer technology(accumulated publications and patents). Colors indicate the application area (PH, EO, EV) and the thickness of the line indicates whether it shows raw data (thin) or the result of fitting the NLS model (thick).

We observe the same phenomenon regarding the starting point of time intervals: both patent and publication data corresponding to this technology show the same starting point, according to our methodology. In this case, however, the pattern shown by the areas under analysis is significantly different for each data source. While the EV applications remain below PH and EO in patent data, research in EV applications is clearly outpacing EO and

shows signs of surpassing the PH area, as can be seen in the *b* parameter values shown in Table 1. EO and EV areas seem to be catching the attention of the scientific community at a fast pace when compared with PH, by looking at their *b* values. This was not the case in SST technology. However, PH applications of SiC based-transformers are the most frequent, according to our data.

Table 1 shows the values of *b* for the NLS models built for each technology, area and data source:

Table 1. Values of b for each technology, area and data source.

Technology - Area	Data source	*b* parameter
SST-PH	Development (patents)	2.15
SST-EO	Development (patents)	1.97
SST-EV	Development (patents)	2.79
SST-PH	Research (scientific pub.)	2.39
SST-EO	Research (scientific pub.)	2.33
SST-EV	Research (scientific pub.)	2.31
SIC-PH	Development (patents)	3.54
SIC-EO	Development (patents)	3.07
SIC-EV	Development (patents)	3.38
SIC-PH	Research (scientific pub.)	2.57
SIC-EO	Research (scientific pub.)	3.58
SIC-EV	Research (scientific pub.)	3.91

The results show that EV applications have the highest average *b* coefficient (3.09), while PH and EO applications show similar *b* values (2.66 and 2.73, respectively).

5. Discussion and conclusions

A remarkable conclusion of this study is that the ideas of the linear model of innovation (Godin, 2006) fail to describe the behavior of SST and SiC transformer technologies: both research and development seem to be taking place simultaneously, according to our data. Research in these technologies is eminently applied science, where the boundaries between science and development become more porous (Kline, 1985), this could be an explanation of

the phenomenon we have observed, but for our purposes, this trait of the technologies hinders the detection of early signs of emergence coming from the scientific world.

A head to head comparation between the two renewable energy sources under study suggests that the PH area will probably outpace the EO area in both technologies (SST and SiC-based transformers). The trends in scientific research in SiC-based transformers, however, suggest that the attention of the scientific community might be shifting from PH in this technology.

Perhaps the most interesting pattern can be found in the EV applications of both SST and SiC transformers. Data corresponding to EV area persistently shows a smaller yearly amount of research/patenting activity taking place when compared with the rest of the areas (with a single exception) but at the same time the higher average *b* parameter is found in this area. According to our approach, this suggests that EV applications might be closer to the life cycle's exponential growth stage than PH and EO areas, so it seems reasonable to expect a comparatively faster development of both scientific and development activity in this field, when compared with PH and EO. This conclusion is reinforced, from our point of view, by the data presented in Figure 2 (right), which points at both a strong presence and exponential growth pattern in the academic activity related to SiC transformer applications in EV area. Considering that a substantial amount of research related with new semiconductor materials falls into the realm of basic science, technology forecasting efforts in this area should probably keep an eye on this sample of data, in order to look for early signals of emergence.

We hope that the results and the conclusions presented in this study will be useful for decision making in the field of renewable energies and the technologies related to electric vehicles, particularly for those professionals involved in technology forecasting practices.

References

Abu-Siada, A., Budiri, J., & Abdou, A. (2018). Solid State Transformers Topologies, Controllers, and Applications: State-of-the-Art Literature Review. *Electronics*, *7*(11), 298. https://doi.org/10.3390/electronics7110298

Bhattacharya, S. (2017). Smart Transformers Will Make the Grid Cleaner and More Flexible. Retrieved January 21, 2020, from https://spectrum.ieee.org/energy/renewables/smart-transformers-will-make-the-grid-cleaner-and-more-flexible

Godin, B. (2006). The Linear Model of Innovation: The Historical Construction of an Analytical Framework. *Science, Technology & Human Values*, *31*(6), 639–667. https://doi.org/10.1177/0162243906291865

Hastie, T. J. (2017). *Statistical models in S*. (T. J. Hastie & J. M. Chambers, Eds.). Routledge.

Kline, S. (1985). Innovation is not a linear process. *Research Management*. Retrieved from http://www.ec.unipg.it/ez_new/index.php/ita/content/download/7711/35914/file/FILE 3_ Kline_ Innovation is not a linear process.pdf

Kucharavy, D., & De Guio, R. (2015). Application of Logistic Growth Curve. *Procedia Engineering*, *131*, 280–290. https://doi.org/10.1016/J.PROENG.2015.12.390

Roberts, D. (2018). Renewable energy threatens to overwhelm the grid. Here's how it can adapt. Retrieved January 21, 2020, from https://www.vox.com/energy-and-environment/2018/11/30/17868620/renewable-energy-power-grid-architecture

Ronanki, D., Kelkar, A., & Williamson, S. S. (2019). Extreme Fast Charging Technology—Prospects to Enhance Sustainable Electric Transportation. *Energies*, *12*(19), 3721. https://doi.org/10.3390/en12193721

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11322

Mining News Data for the Measurement and Prediction of Inflation Expectations

Diana Gabrielyan[1], Jaan Masso[1], Lenno Uusküla[2]
[1]University of Tartu, Tartu, Estonia, [2]Bank of Estonia, Tallinn, Estonia.

Abstract

In this paper we use high frequency multidimensional textual news data and propose an index of inflation news. We utilize the power of text mining and its ability to convert large collections of text from unstructured to structured form for in-depth quantitative analysis of online news data. The significant relationship between the household's infla-tion expectations and news topics is documented and the forecasting performance of news-based indices is evaluated for different horizons and model variations. Results sug-gest that with optimal number of topics a machine learning model is able to forecast the inflation expectations with greater accuracy than the simple autoregressive models. Addi-tional results from forecasting headline inflation indicate that the overall forecasting accu-racy is at a good level. Findings in this paper support the view in the literature that the news are good indicators of inflation and are able to capture inflation expectations well.

Keywords: *inflation; inflation expectations; news data; natural language processing; topic modelling.*

1. Introduction

Household surveys of inflation often indicate that the perception of the current inflation differs substantially from the actual values of inflation. Similarly, expectations about the future expectations differ strongly from the surveys of professional forecasters and the implied inflation rates of financial markets (for evidence see e-g- Coibion et al. 2018). Potential reason for the difference is that households and firms obtain only very partial information while doing everyday purchases and aggregating the information is very costly. Imperfect information in turn affects adversely the formation of expectations. Subjective inflation nowcasts and expectations are built through personal experiences, prior memories of inflation, and various other sources of information. One primary source of information is public media and it is well established that consumers rely largely on it when thinking about overall price changes (Blinder and Krueger 2004, Curtin 2007). Media covers a lot of news on prices and price developments.

In this paper we explore online news as novel data source for measuring inflation perception and forecasting inflation expectations by utilizing the power of text mining and its ability to convert large collections of text from unstructured to structured form. We propose a novel index of inflation news that provides a real-time indication of the price developments. Such index of inflation news captures and summarizes well the information used in the formation of expectations[1]. Available survey-based inflation expectations have low frequency and the high-frequency market-based forecasts involve risk premia and may be uncertain[2]. Our main contribution is therefore using the novel source of information to prove that online news can provide a real-time and accurate indication of consumer's expectations on inflation.

Machine learning methods are considered to be very promising avenue for academic and applied research. Although its applications are already actively used in many disciplines and research areas, it is still relatively new to economics. One modern strand of machine learning is text mining – the computational approach to processing and summarizing large amounts of text, which would be far more difficult to read, even impossible, for any single person. Extracting information from novel sources of data, such as social media (e.g. Twitter, Google) or public media (e.g. online news, communication reports) allows analysis and

[1] As Nimark and Pitschner (2018) note, since no agent has resources to monitor all events potentially relevant for his decisions, news are preferred delegates for information choice to monitor the world on their behalf. And since news mainly reports selection of events, typically major ones, coverage becomes more homogenous across different outlets.

[2] Market-based expectations are available daily but include risk premia. Survey-based expectations are published monthly. For example, for the United Kingdom, the quarterly Consumer trends data are typically published around 90 days after the end of the quarter. See https://www.ons.gov.uk/economy/nationalaccounts/satelliteaccounts/bulletins/consumertrends/apriltojune2019

different kind of understanding of economics relationships, e.g. consumer behaviour, therefore contributing to policy making and forecasting. See for example, Tuhkuri (2016), D'Amuri and Marcucci (2017), Yu et al (2018), Nyman et al (2015).

Another contribution of this work is to forecast the inflation in real-time using machine learning methods. The importance of inflation forecasting for rational decision making is well established in the literature along with the common knowledge that improving upon simple models is quite challenging. According to Medeiros et al (2019), most of this literature however ignores the recent machine learning advances. In their work they show that with machine leaning and data-rich models improving inflation forecasts is possible. Their LASSO and Random Forest models are able to produce more accurate forecasts than the standard benchmark models, e.g. autoregressive models. Similarly, Garcia, Medeiros and Vasconcelos (2017) find that high dimensional models perform very well in inflation forecasting in data rich environments. Our findings from LASSO regressions support these findings: for inflation expectations the short-term forecast errors are smaller than those of the autoregressive models. The analysis also identifies the optimal number of news topics for predicting up to five quarters ahead inflation expectations to be either four or five suggesting that the LASSO regression using optimal number of topics and best value of regularization parameter results in simpler model, which doesn't compromise the model performance. These results are, however, not robust for longer forecasting horizons and for different values of the regularization parameter. In additional results, when forecasting headline inflation, we find that the LASSO models fail to improve upon the benchmark models but demonstrate similar forecasting accuracy.

The rest of the paper is organized as follows. Section 2 describes the data sources and methodology. Section 3 and 4 provide results and an application in forecasting respectively. Section 5 concludes.

2. Data and Methodology

For official statistics, we use the Bank of England Inflation Attitude Survey data and actual UK inflation statstics. Our novel inflation news indicator is built from the article data of one of the UK leading newspaper's , Guardian, business section over the last 15 years. The choice of the news outlet is due relevance to our research in terms of content and readership, as well as the availability of open source data. As such, we chose Guardian news data for our analysis. Any news in Guardian is public and readable by anyone by default. Overall, we collected around 20,000 documents and 32 million terms from January 2004 to January 2019, which is sufficient amount of data to conduct our analysis. We only fetch articles from the business section, since this is the most relevant section for economic topics in general. In addition, articles were also filtered based on subjectively chosen key-words, which in our

opinion are relevant to inflation expectations topic. Namely, they are price, price increase, expensive, cheaper, cost, expense, bill, payment, oil, petrol, gas, diesel. The data comes in unstructured form, that is, the data is in a text form and does not have a given structure. Overall, our news corpus consists of around 100000 English language articles with well above 20 million words from January 2004 to January 2019, which is sufficient amount of data to conduct our analysis. However, the amount of data, also makes statistical computations challenge. We therefore apply data pre-processing steps suggested by Bholat and co-authors (2015) at the same time adding more steps and more developed methods. We use the text mining's bag of word approach in the text, which means all words are analysed as a single token and their structure, grammar or part of lexicon does not matter. Pre-processing results in a document term matrix, which includes all occurrences of the words in the corpus and their respective frequencies. At this step, the dimensionality of the corpus is reduced, and we get more understandable results. Frequency counts of the top 31 words in their stemmed form, that is the number of times those words appear in the final sample, are plotted in Figure 1.

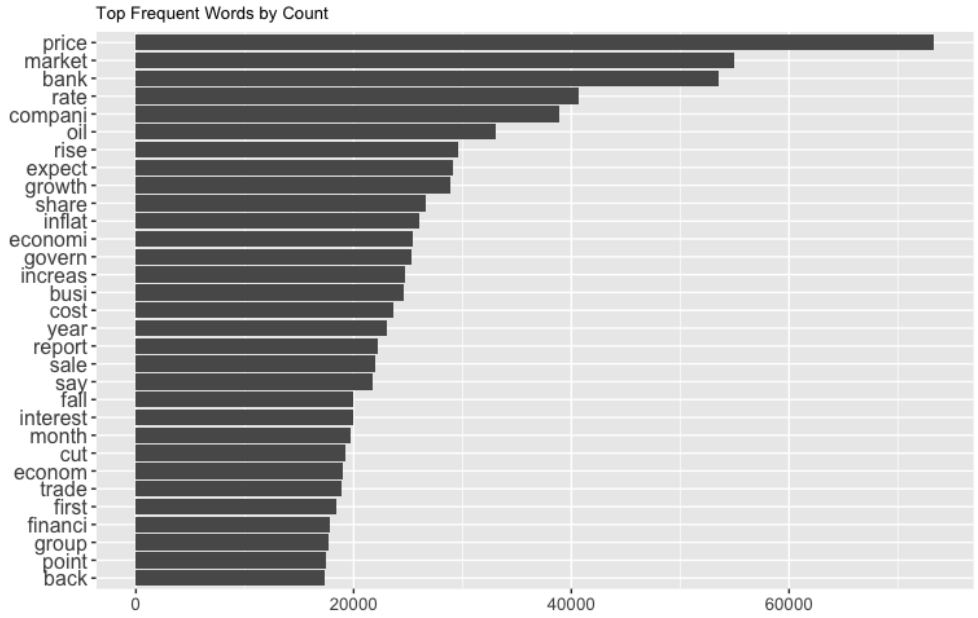

Figure 1. Top frequent words and their counts. The words are presented in stemmed form.

To proceed to building the index, we proceed with topic modelling. Since any document can be assigned to several topics at a time, the probability distribution across topics for each

document is therefore needed. Latent Dirichlet Allocation (LDA) [3] is a statistical model that identifies each document as a mixture of topics (related to multiple topics) and attributes each word to one of the document's topics, therefore, clustering words into topics. With LDA method, it is possible to derive their probability distribution by assigning probabilities to each word and document. Assigning words and documents to multiple topics also has the advantage of semantic flexibility (ex. the word 'rate' can relate both to inflation and unemployment topic). Thorstrud (2018) notes that LDA shares many features with Gaussian factor models, with the difference being that factors here are topics and are fed through a multinomial likelihood. In LDA, each document is given a probability distribution and for each word in each document, a topic assignment is made.

3. Results

For each document within a day, five most popular words are identified, and their daily frequency is calculated. This allows counting also the frequency of each topic for a given day. At this step, our results of topic decompositions and distribution are used to build the new high frequency index that will capture the intensity of inflation expectations. The index is built for every day, that is, we build daily time series using Guardian's business articles for each day. To do so, we first collect together all articles for a given day into one document, grouping them into one plain text for each day. Next, based on the first ten most frequent words in each topic the article's daily frequency is calculated. In other words, the frequency is calculated for the given day as the raw count of frequencies with which the most common words in each topic appear in that day. The news volume *I(t)* of given topic *z* is given by

$$I_z(\mathrm{t}) = \sum_{d \in I(t)} \sum_w N(d, w, z), \tag{2}$$

where $N(d, w, z)$ is the frequency with which the word *w* tagged with topic z appears in document d. These time series $I_z(\mathrm{t})$ are measures of volume, that is, they measure the intensity of given topic for given time period, that is for given day.

We find that some of index series are non-stationary and consequently transform them to stationary series by differencing. Augmented Dickey Fuller test is used to determine the presence of unit root and hence understand if the series are stationary or not. As such, some of the indices are evaluated as non-stationary and are transformed to by differencing.

[3] Detailed description of the LDA approach is provided in Blei, Ng and Jordan (2003).

4. Application in Forecasting

The first task is to filter information from the list of variables and select more relevant components. It is highly inefficient to use all the topic indices for predicting in such a rich dataset, as some of the regressors may be imparting redundant information. Therefore, number of topics *N* is too high and there is a definite multicollinearity present among the topic indices, as can also be observed from Figure 4. To reduce dimensionality and tackle the issue of multicollinearity[4], we use another machine learning method for variable selection. LASSO (Least Absolute Shrinkage and Selection Operator) method automates variable selection by reducing the coefficients of some features to zero, while keeping those that have the most impact on the dependent variable. LASSO's main goal is finding β that minimizes (3) with constraint $\sum_{j=1}^{p}|\beta_i| \leq t$.

$$\sum_{i=1}^{M} (y_i - \sum_{j=1}^{p} \beta_i x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_i| \quad (3)$$

λ is the shrinkage parameter and controls the strength of penalty finding the model with the smallest number of predictors that also gives a good accuracy. Therefore, the number of variables to be removed is decided by the shrinkage parameter λ, which is chosen using cross validation. Once the topic indices are selected, we forecast the inflation expectations by building a model using a direct forecast approach as given in equation (4).

$$\pi_t^{t+h} = \alpha + a * \pi_{t-1}^{t-1+h-1} + \sum_{n=1}^{N} b_n * x_{n,t-1} + u_t, \quad (4)$$

where π_t^{t+h} is the inflation (expectations) for the next *h* quarters at period *t* and $\pi_{t-1}^{t-1+h-1}$ the lagged value for the same horizon. *N* is the number of indices built from news data, b_n are vectors of unknown parameters, $x_{n,t}$ are the lagged indices and u_t is the forecasting error. We call the Equation (3) a news-based model (NBM). It is common practice to fit a model using training data, and then to evaluate its performance on a test data set. Forecast horizon *h* is also the length of the out-of-sample period (i.e. fitted values on the training set) and will be varied from 1 to 12 to compare the forecasts at different horizons and find the 'optimal' horizon defined by the lowest forecasting error. Since all of the data in this analysis is quarterly, *h* is measured in quarters. For benchmarking we use naïve AR (1) model on inflation expectations and compare the root mean squared errors (RMSE).

Table 1 reports the normalized results of estimating (3) and an AR (2) with different forecast horizons relative to simple AR (1) model. The first column of the table shows the forecast horizon, the second column (*n_var*) shows the number of variables (topics) selected by LASSO regression, and the last two columns show the root mean squared errors (RMSE) for each of the applied models. It can be seen that generally, the RMSEs are small, varying from

[4] LASSO is very robust against multicollinearity, see Friedmanet al. (2001).

0.02 to 0.76), while the forecast errors are the lowest when forecasting the next one or two period expectations using the news data. In this case the LASSO model outperforms both the naïve AR (1) and AR (2) forecasts in terms of accuracy.

Table 1. RMSEs of h-period inflation expectations forecasts using LASSO and AR (2) models. Errors are normalized relative to AR (1) benchmark.

h	n_min	RMSE_LASSO_MIN	RMSE_AR2
1	5	0.6	6
2	5	0.7	1.9
3	6	0.9	1.8
4	5	0.8	1.8
5	5	0.8	1.8
6	5	1	1.6
7	5	1.1	1.7
8	5	1	2.1
9	5	1.2	2.9
10	4	1.6	1.9
11	3	1.3	1.8
12	3	1.5	1.8

Several interesting observations can be made from Table 1. Firstly, LASSO models select different number of topics that are relevant for inflation expectations prediction for different forecast horizons. Out of our fifty topics compiled by the LDA method, LASSO selects three to six topics depending on the forecast horizon. Lagged value of the inflation expectations is always included among selected regressors and is always significant. The adjusted R-squared statistic is informative and for some horizons is as high as 70%. Thus, the selected news topic, as well as the past values of inflation expectations explain a relatively large fraction of the variation in the household's inflation expectations. One to two quarters ahead expectations can be forecasted with five topic indices as regressors, while the longer forecasts of eleven and twelve quarters can be forecasted with the best accuracy when only three

relevant topics are employed in the regression. It can also be observed that the longer the forecast horizon, the lower the forecast accuracy, which is intuitive.

These results were not robust when controlling and comparing different values of regularization parameter in the LASSO regression. There are different ways to choose the optimal value of lambda by cross-validation. Our main results in Table 1, where based on the smallest value of lambda from the cross-validation results. Table 2 compares the accuracy obtained with LASSO regression using different values of lambda shrinkage parameters against the benchmark autoregressive models. First column is the forecast horizon, while following 3 columns report the number of regressors selected by LASSO for different values of lambda. Among selected topics for all three variations of lambda, first lag of inflation expectations is selected. Column RMSE_LASSO_MIN uses the value of lambda that is equal to the minimum value of lambda chosen by cross-validation, while column RMSE_LASSO_LSE is based on the model where lambda is within one standard error. Column RMSE_LASSO_BIC is based on the lambda which is chosen using information criterion. Last two columns show the errors for benchmark AR (1) model AR (2) model, normalized relative to AR (1). Given the sparsity across normalized errors for different forecast horizons, as well as in the number of topics selected by LASSO, it can be noted that LASSO models other than that based on its minimum value are less accurate and fail to outperform the naïve models.

Table 2. RMSEs of h-period inflation expectations forecasts using different values of lambda in LASSO model, as well as AR (1) and AR (2) models. All values are normalized relative to AR (1) benchmark.

h	n_min	n_lse	n_bic	RMSE_LASSO_MIN	RMSE_LASSO_LSE	RMSE_LASSO_BIC	RMSE_AR1	RMSE_AR2
1	5	2	4	0.6	3.4	0.4	1	6
2	5	2	4	0.7	1.8	0.7	1	1.9
3	6	2	48	0.9	3.5	5.9	1	1.8
4	5	2	50	0.8	3.7	6.1	1	1.8
5	5	2	4	0.8	4.5	0.7	1	1.8
6	5	2	44	1	4.6	7.1	1	1.6
7	5	2	47	1.1	5.1	7.4	1	1.7
8	5	2	41	1	5	7.5	1	2.1
9	5	2	41	1.2	4.9	6.8	1	2.9
10	4	3	2	1.6	1.3	1.5	1	1.9
11	3	2	2	1.3	1.3	1.5	1	1.8
12	3	2	2	1.5	1.5	1.6	1	1.8

The model obtained from RMSE_LASSO_LSE includes less topics but shows poor forecasting performance. Similarly, the model from RMSE_LASSO_BIC includes even more predictors, particularly in the intermediate horizons, however, shows even worse performance. In the shorter forecasting horizons, the number of chosen topics is four, which is closer to five from the minimum lambda model, and the forecast accuracy improves. These analyses demonstrate that the optimal number of topics to predict inflation expectations up to five quarters ahead are between four and five. This also suggests that the LASSO regression, using minimum lambda as the best lambda, results to simpler model without compromising much the model performance on the test data.

It is also of interest to look how the same news data and model can be used to predict the headline inflation. We computed forecast errors for different horizons and models compared to benchmark AR (1) for annual rate of inflation and its quarterly rate. Results, not included in this chapter, but available from authors upon request suggest that while the LASSO model built using pre-selected news topics does not outperform the benchmark models, it can however be used as a forecasting model with similar forecast accuracy as those naïve models. This means that the model obtained with LASSO regression does at least as good a job fitting the information in the data as the more complicated one.

References

Alan S. Blinder & Alan B. Krueger . What Does the Public Know about Eco-nomic Policy, and How Does It Know It? *Brookings Papers on Economic Activity, Economic Studies Program, The Brookings Institution* 35(1), 327-397

Francesco D'Amuri & Juri Marcucci (2017): The predictive power of Google searches in forecasting US unemployment*, International Journal of Forecasting*, Vol. 33, No. 4, pages 801-816.

David M. Blei, Andrew Y. Ng & Michael I. Jordan (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research* 3, 993–1022.

Erik Cambria & Bebo White: Jumping NLP Curves (2014). *A Review of Natural Language Processing Research, proceedings of Research Review Article IEEE Computational intelligence magazine*, 9, pp. 48.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. Vol. 1. Springer series in statistics Springer, Berlin.

Kristin J. Forbes, Lewis Kirkham & Konstantinos Theodoridis (2017). A Trendy Ap-proach to UK Inflation Dynamics. *Bank of England Working Paper* 49.

Neil Gerstein, Bart Hobijn, Fernanda Nechio & Adam H. Shapiro (2019). The Brexit price Strike. *FRBSF Economic Letter*, Federal Bank of San Francisco.

Leif A. Thorsrud (2018). Words are the new numbers: A newsy coincident index of business cycles*, Journal of Business & Economic Statistics*.

Marcelo C. Medeiros, Gabriel F. R. Vasconcelos, Álvaro Veiga & Eduardo Zil-berman (2019). Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods, *Journal of Business & Economic Statistics*.

Kristoffer P. Nimark and Stefan Pitschner (2018). News Media and Delegated Infor-mation Choice. CEPR Discussion Papers 11323, *C.E.P.R. Discussion Papers*.

Joonas Tuhkuri: Forecasting Unemployment with Google Searches (2016). *ETLA Working Papers*, No 35.

Yu L, Zhao Y, Tang L and Yang (2018). Online big data-driven oil consumption fore-casting with Google trends, *International Journal of Forecasting.*

Rickard Nyman, David Gregory, Sujit Kapadia, Paul Ormerod, David Tuckett & Robert Smith (2015). News and narratives in financial systems: exploiting big data for systemic risk assessment, mimeo.

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11359

Citizens' attention in Madrid City through the study of personalized records

Pilar Rey del Castillo
Instituto de Estudios Fiscales, Ministry of Finance, Spain.

Abstract

The datification of our daily lives in the Big Data era is producing a huge amount of information about processes and activities that were previously invisible or at least difficult to grasp, leading to new opportunities and challenges for analysis.

Examples of some data available are the tens of million of Personalized Attention Records that can be downloaded from the open data portal offered by the local government of Madrid City. These records become a sort of counterpart from the call receiver's perspective of the Call Detail Records produced by telecom providers. They are stored as a result of a front office tool retaining some information from a range of different communication channels to manage the interaction with users.

The paper explores the data contained on these Personalized Attention Records to help improve customer attention services. It emphasizes the study of the topics that concern the citizens and the different channels dealing with the services, using Natural Language Processing and other tools.

Keywords: *Big Data; Call Records; Natural Language Processing.*

1. Introduction

The Personalized Attention Records (PARs) of Linea Madrid are between the datasets made available by the local government of Madrid City in its open data portal https://datos.madrid.es/portal/site/egob. These records may be considered as a sort of counterpart from the call receiver's perspective of the Call Detail Records produced by telecom providers. The source for the PARs is the Customer Relationship Management (Buttle and Maklan, 2015), a front-end tool offering an interest oriented management solution. It gathers the data from different communications channels: the 26 citizen attention offices distributed by borough, the 010 phone number, the website chat, the Facebook account and the Twitter account @lineamadrid. The volume of the downloaded information, more than 44 million of records from 2014, cannot be processed using conventional statistical software and requires procedures specially developed for this purpose. Apache Spark (Zaharia et al., 2016), an open source analytics engine for Big Data processing has been used for the first steps of collecting and pre-processing data. Besides this, Python software (Van Rossum & Drake, 2009) and Scikit-learn (Pedregosa et al., 2011), a free software machine learning library for the Python programming language have been used for further calculations and analysis.

Each record contains a number of variables that have been changing through time. Some of these variables, such as the responsible worker or whether the issue has been addressed to another instance, are mostly interesting for administrative purposes. But there are other variables whose analysis may help to improve customer attention services. Besides the reception and register date, this paper focuses on the study of variables remaining through time, such as the topics that concern the citizens and the different channels dealing with the services.

Table 1. Examples of topic description variables in Personalized Attention Records.

Tipo 1	**Tipo 2**	**Tipo 3**	**Tipo 4**
Información general	Administración Pública	Administración estatal	
Movilidad	Madrid Central	Alta personas	
Identificación electrónica	Acceso a Carpeta Ciudadano	Alta	
Tasas e impuestos	IBI	Consulta/Información	Voluntaria
Cita Previa	Cita Previa	Asignar cita previa	
Movilidad	Multas	Pago con tarjeta	Voluntario
Padrón municipal	Justificantes empadronamiento	Volante empadronamiento	
Registro	Registro	Anotación	
Avisos	Avisos	Alta/Reiteración	

There is a specific variable reflecting the channel while the topic is spread out over four variables (tipo1 to tipo4). The target of using four variables to collect the topic seems to be obtaining a hierarchical description allowing for detailed information. But in practice the data have been filled in in various ways as can be seen in Table 1.

Daily indicators of the number of requests or questions received by channel and topic will be computed to provide an idea of the manner in which citizens' attention is managed by municipality services. For this purpose, the topic description variables need to be previously treated by Natural Language Processing tools.

The remainder of this paper is organized as follows. Next section describes the first steps of processing the records, making them homogeneous and obtaining a realistic classification of topics; section 3 presents and analyses the results obtained for the period between January 2014 and March 2020; and finally, a number of remarks and conclusions are presented in Section 4.

2. Processing of the Personalized Attention Records

The datasets including the records registered in the month are made available in the Madrid City open data portal after the end of each month. These files include the information of a number of variables varying in time making around 700 000 data points for each year and each variable.

The first step consists always on detecting and correcting possible logical inconsistencies in the data. For instance, for each new dataset, changes on date formats and variables appearing or disappearing are frequently found and must be previously detected to allow for subsequent homogeneous treatments. Likewise, once within the file, common spelling errors in string variables that have a fixed number of categories can be detected and corrected.

After the previous corrections, the selected data reflect the reception date, the channel receiving the request and the topic description variables. These last four variables (tipo1 to tipo4) must then be treated to obtain practical categories for classifying the topics which the requests refer to. For this purpose, some steps of Natural Language Processing (Jurafsky & Martin, 2008) have been carried out:

- Concatenation of variables tipo1 to tipo4 into one topic string, and conversion into lower case letters.
- Tokenization of the topic string into words.
- Elimination of duplicate words.
- Elimination of Spanish stop words (most common words that are filtered out).

These steps result in a list of significant words reflecting the topic for each record. A wordcloud is shown in Figure 1, which consists of a visual representation of the words in the topic descriptions where the size of each word is proportional to its frequency (Halvey and Keane, 2007).

Figure 1. Wordcloud for the words reflecting the topic for the Personalized Attention Records.

The first attempt to obtain a classification for the topics was done using Latent Dirichlet Allocation (Blei et al., 2003), a method for unsupervised classification that try to identify the topics best describing a set of documents. It builds a probabilistic model where each topic is characterized by a distribution over words, and documents are assigned probabilities to come from each topic. In the case of the PARs, the list of words for each record is too short, producing not interesting topics.

The next step has been finding the most frequent lists of one or more words by means of the Frequent Pattern Growth Algorithm, an efficient method of finding association rules (Witten et al., 2016). A classification on 15 classes has been established using these lists: *Bicycle, Culture and Sport, Elderlies, Electronic Signing, Environment, Employment, Foreigners, Health, Central Madrid*[1]*, Mobility, Other, Penalties, Register, Taxes,* and *Urban Planning*. In this way, each record is classified in one of the categories depending on the words in its topic description.

[1] New big traffic-restricted area

3. Analysis of the results

For each record the data include now the date, the channel receiving the request and the topic it refers to. A first question to raise is about the relationship between channel and topic, whether there is any kind of association between these two nominal variables. The most popular measure of relationship between this type of variables is Cramer's V (Cramér, 1946) that takes values between 0 and 1, with values closer to 1 indicating a greater association. In this case its value is 0.20, reflecting a low-medium level of relationship.

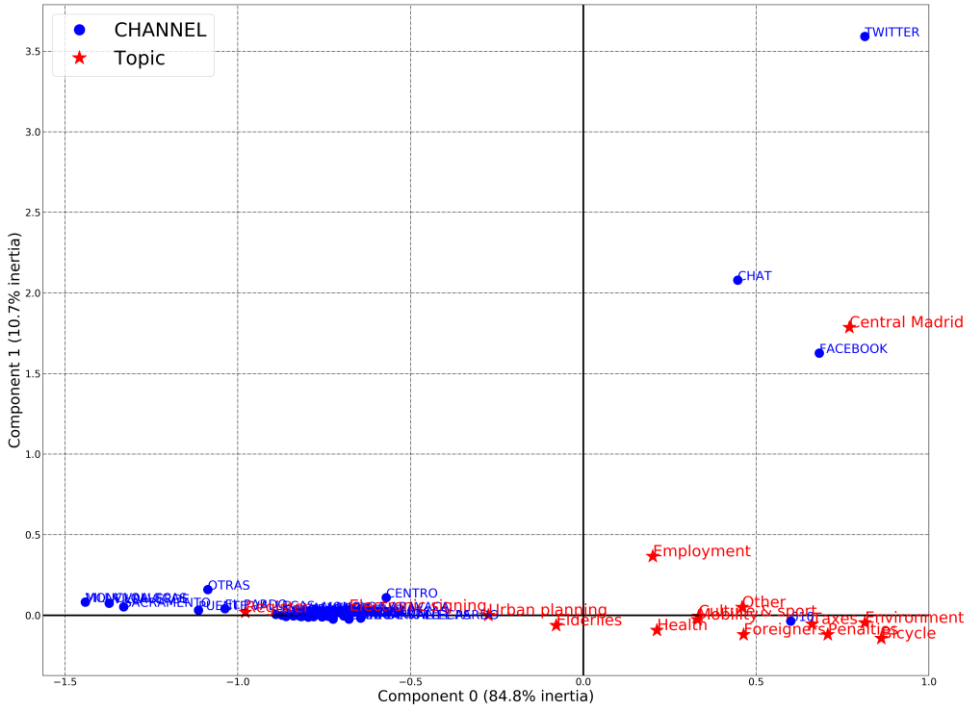

Figure 2. Principal coordinates for Channel and Topic categories.

Para seguir leyendo, inicie el proceso de compra, click aquí