Data Mining and Social Sciences in Spain;

Conclusions on the state-of-the-art methods for political forecasting with Twitter

Xavier Arque Bota

xbota@uoc.edu

Submitted in the partial fulfillment of the requirements for the degree of

MASTER GEICO

[Gestió Estratégica de la Informació i el Coneixement]

University – UOC

Director TFM: Dr. Josep Cobarsí Morales

Data Mining and Social Sciences in Spain;

Conclusions on the state-of-the-art methods for political forecasting with Twitter

Xavier Arque Bota

Abstract

For any social scientist to have socioeconomic quality information is crucial and because the content of Twitter messages (or any SNS) plausibly reflects the offline political landscape it's important for social scientist to be able to tap this information. But in a digital world, with data being generated massively and in an unstructured format, the classical tools to capture information for socio-political analysis have been partly outdated. In this work I pretend to introduce Social Science researchers to the opportunities and challenges of using Twitter data to analyze political trends. I review the literature and the researchers profile, I discuss the use of microblogging message content as a valid indicator of Spanish political sentiment, and assess the relations between Social Sciences and Computer Science. The main contributions of this paper are two-folds: First; traditional polls accuracy is still higher. Although the gap is narrowing in some areas, like big tendencies detection, there is a need for better tools and better theoretical frameworks. Second; the results display a research landscape occupy by Computer Science researchers and few Social Science researchers, mostly working in the USA. This is bad because it atomizes the research. Looks like most of the Social Scientists, the ones that have to create the theoretical framework, don't have the tools or the knowledge to work with Big Data.

Keywords

Twitter, politics, Social Sciences, forecast, meta-review

Introduction

1 Justification

Social media on-line services have spread throughout the world in just a few years and online interaction via Facebook, Instagram, Twitter, etc have become part of the daily life for a demographically diverse population of billions of people worldwide. But does this huge amount of users make Social Media data a valid indicator of socio-political behavior? And, digital data tracking what people does and the time and place of their actions is feasible but, may it aid social researchers to better understand and forecast socio-political trends and population attitudes?

So far, the social networks created among Facebook users, Twitter users or any other Web 2.0 social network site (SNS) users cannot be equated with offline networks based primarily on face-to-face interaction. The way people communicate in the real-world versus the way people communicate using Internet or any digital-network are getting closer but there is still a big gap in the amount and quality of the information transmitted due to a lack of physical interaction (eye contact, handshakes, hugs, etc), traditional institutions (dressing codes, languages, appearance, etc) and because of the different demographics involved, that is to say, not all the people are on-line neither the stratification in society or in the social networks is proportional (Mislove, Lehmann, Ahn, Onnela, & Rosenquist, 2011) (Chung & Mustafaraj, 2011)

Nevertheless, recent studies have compared the volume and direction of messages re-tweets and @ mentions among Twitter followers with the same users' offline friends and found out a close correspondence (Xie et al., 2012) And because real-world events have an impact on online systems, the trails left by users on such systems have been used to evaluate people response to different socio-economical aspects of their live and used as well for events detection. Blog posts have been correlated

with book sales (Gruhl, Guha, Kumar, Novak, & Tomkins, 2005) with movie gross incomes (Mishne & Glance, 2005) or with collective action and social movement mobilization (Gonzalez-Bailon, Borge-Holthoefer, Rivero, & Moreno, 2011) and Twitter data have been used to measure unemployment, spread of disease, consumer confidence, social mood, investor sentiment, financial market behavior, brand loyalty, etc . (Bollen, Pepe, & Mao, 2009; DiGrazia, McKelvey, Bollen, & Rojas, 2013; He, Zha, & Li, 2013; Liu, 2012; Pang & Lee, 2008; Tan et al., 2014)

Researchers have also used changes in the distribution of user-generated content not only to explain or review different aspects, attitudes and facts of our society but to try to predict them. DiGracia (DiGrazia et al., 2013) showed that local US Election outcomes were positively correlated with the numbers of times that Republicans had been mentioned in tweets. O'Connor (O'Connor, Balasubramanyan, Routledge, & Smith, 2010) correlated tweets with several public opinion time series; Tumasjan (Tumasjan, Sprenger, Sandner, & Welpe, 2011) predicted the outcome of German elections; Bollen (Bollen et al., 2009) forecast the stock market directions through twitter user's moods; and Lampos and Cristianini (Lampos, De Bie, & Cristianini, 2010) predicted the evolution of flu pandemics.

Among these networks, Twitter, with over 310 million active users who collectively generate more that 550 million tweets each day[1], stands out as by far the largest and most comprehensive publicly accessible source of online data on social interaction. In Social Sciences research, Twitter has earned the focus of extensive research largely due to its openness in sharing public data, its real-time nature and the easiness of data collection thanks to its application programing interfaces (APIs) that can be used easily to collect a wealth of social data through out tweets and Tweeter users. Regardless of their content and intended use, tweets often convey pertinent information about their authors mood status. As such, tweets can be regarded as temporally-authentic microscopic instantiations of public mood state (O'Connor et al., 2010) and patterns found in digital trace data are increasingly used as evidence of social phenomena. Following this line of research, J. Soler (Soler, Cuartero, & Roblizo, 2012) states "that the references to

1 https://about.twitter.com/company *All numbers approximate as of March 31, 2016.*

the different political parties correlate, significantly, with the votes of the electors (...) and this is an indicator that Twitter may be used by social researchers as a tool, among others, to predict future results of the elections. Of course, with due caution because the measured data correspond to distinct actions, so obviously, much more research and studies should be done in this field".

Nevertheless, recent papers (Gayo-Avello, 2012b) concluded that predictive claims are exaggerated and Gayo-Avello et al (Gayo-Avello, Metaxas, & Mustafaraj, 2011) found no evidence of correlation between the analysis results and the electoral outcomes. Ceron-Curini (Ceron, Curini, & Iacus, 2014) proved that electoral forecasts accuracy may change depending on the techniques of data-mining and sentiment analysis used and the country or region selected and other researchers have suggested that even the social media sampling strategy may define the discovery or not of dynamic processes like diffusion (Choudhury et al., 2010) Another article (Conover, Ratkiewicz, & Francisco, 2011) aiming to assess ideological polarization on the twitter-sphere has reported opposite findings depending on whether the network was reconstructed using RTs or @mentions. Ming FaiWong (Wong, Sen, & Chiang, 2012) raised some doubts on the predictability of box office performance from Twitter data; and, in a similar way, Jungherr (Jurgens, Jungherr, & Schoen, 2011) and (O'Connor et al., 2010) rebutted that tweet volume and votes are strongly correlated, which was the main thesis by Tumasjan et al. (Tumasjan, Sprenger, Sandner, & Welpe, 2010)

Lastly, positive studies about electoral prediction like (Barbera, 2016; Gayo-Avello et al., 2011) do not claim that political predictions from social data are unfeasible, but that they are much more difficult to obtain than mere tweet volume counting or plain sentiment analysis.

2 Research Goals

Nowadays, thanks to social networks like Twitter, social scientists have more information about communication patterns and people behavior than ever before, and the access to this information is easy and cheap. We also have the tools and the computational power to make sense out of this wealth of

knowledge and to figure out what those patterns reveal. But so far there is no agreement about the real value and accuracy of this knowledge; even worse, the arguments, models and discussions are taking place between researchers mostly not related with the Social Sciences. Thus, as Sandra González points out (González-Bailón, 2013 p-9) “the models that are being applied to social systems have been developed by mathematicians, physicists, or computer scientists to understand the behavior of other systems that have no agency (...). The models we build about social systems with the help of Big Data should be consistent with what we know about human actors and their behavior”.

Twitter data opens up transformative possibilities for both descriptive and analytical Social Studies, but without the automated data management and coding tools developed by computer scientists the analysis of massive unstructured data will remain beyond the reach of most social scientists (Mejova, Weber, & Macy, 2015a) Social scientists have to embrace Big Data and gain literacy in the research that makes it possible. This is important not only to empower themselves when dealing with massive datasets but also to make them feel skilled when joining research teams with researchers beyond traditional disciplinary boundaries. We can not leave the field to disciplines that are much better at building powerful telescopes that at knowing where to point them (Lazer et al., 2009)(Golder & Macy, 2014)

3 Definitions

Some of the concepts and words used are important so a small set of definitions may be helpful.

A social network is a social structure comprising of persons or organizations, which usually are represented as nodes, together with social relations, which correspond to the links among nodes. The social relation could be both explicit, such as kinship and classmates, and implicit, for example friendship and common interest (Yu & Kak, 2012)

Social media comprise platforms to create and exchange user-generated content. Social media are different from traditional media in that almost anyone can publish and access information inexpensively. In contrast, traditional media (which is also referred as old media or legacy media) requires significant resources to publish contents (Yu & Kak, 2012)

Twitter mining is analyzing Twitter message information to predict, discover or investigate potential causation. It can include analyzing additional information associated with tweets (names, hashtags, keywords, emoticons, etc). Twitter mining also employs the substantial quantitative information (numbers of tweets, re-tweets, likes, favorites, etc.) to try to better understand the phenomena under consideration. Finally, Twitter mining can examine how Twitter tweets, re-tweets, etc., capture and reflect different events or even how Twitter relates to other social and conventional media. (O'Leary, 2015)

Machine learning refers to statistical techniques that use past observations to classify new observations or make predictions about the associated outcomes. These techniques may be useful when data have nonlinear relationships or a large number of variables that interact in a complex system in ways that cannot be modeled by traditional regression-based methods. Applications range from understanding natural human language to detecting which emails are spam (Golder & Macy, 2014). Examples of its application to the aforementioned tasks can be seen in (Pennacchiotti & Popescu, 2011), (Castillo, Mendoza, & Poblete, 2013) or (Conover et al., 2011)

Sentiment analysis, also called opinion mining, is a type of language processing that examine opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, and topics. For example, companies always want to understand the people's opinions about their products and services, and which particular features are popular with certain demographics. It uses a combination of statistical techniques and human-created lexicons to identify the valence and intensity of various emotional states

expressed in a body of text. The reader interested in sentiment analysis should consult the works by (Pang & Lee, 2008) or (Liu, 2012)

R programing language is a language and environment for statistical computing and graphics. It is a project similar to the S language and environment, thus although there are some important differences, much of the code written for S runs unaltered under R. R provides a wide variety of statistical and graphical techniques, and is highly extensible. Is available as Free Software under the terms of the Free Software Foundation's GNU License in source code form. It runs on a wide variety of platforms.

Support Vector Machine are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

Random Forest is a notion of the general technique of random decision forests that are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

API is a set of routines, protocols, and tools for building applications. A good API makes it easier to develop a program by providing all the building blocks, which are then put together by the programmer. As an example, a programmer who develops apps for Android may use an Android API to interact with hardware, like the front camera of an Android-based device.

Libraries that perform Machine learning, sentiment analysis, and topic modeling apart from different statistical features are available in R or in standalone software packages such as University of Waikato's Weka[2].

2 http:// www.cs.waikato.ac.nz/ml/weka)or Stanford's Topic Modeling Toolbox http://nlp.stanford.edu/software/tmt/tmt-0.4/

Motivations

This study purposed to review expert opinions and views on the state-of-the-art of political prediction using datasets from social networks, mainly Twitter and to assess the research being done from Social Science researchers in this field. Therefore the expert views and papers were examined according to the following variables:

- What they think is feasible to do with withs these data (feasible)
- What they think will probably happen (probability)
- What they would like to see happen (desirability)

1 Goals

This study focused on the following global main aspects:

a) Can methods like data-mining, sentiment analysis or similar replace traditional forecasting methods?

b) Is there a need to modify social scientist's capabilities and skills in order to adapt them to new technological environments?

And specifically the study aims to find out an answer to this questions:

1 - Can methods like data-mining, sentiment analysis or similar be use in Spain more successfully then offline polls?

2 – Are Spanish Social Scientist prepare to use those methods?

To achieve this goal, a **bibliographical review** was chosen as a method of study for variable A (what is feasible) and the **Delphi technique** plus a statistical analysis of the researchers involved in this field was chosen as a method of study for variables B – (what is probable) and C – (what is desirable).

2 Sources of Information

The databases used are numerous, as currently the subject of research forces to broaden the spectrum of databases of scientific publications. Among the databases to be used, are:

Books and e-books

- Biblioteca de la UOC.

- Red de Bibliotecas Públicas de Catalunya (Girona, Torroella)

Papers, magazines and e-magazines

Google Scholar - scholar.google.com

Directory of Open Access Journals - https://doaj.org/

Electronic Journals EBSCO - http://ejournals.ebsco.com

- ISI Web of Knowledge http://www.sciencedirect.com

Doctoral thesis

- TDX, Consorci de Biblioteques Universitàries de Catalunya –CBUC– http://www.tdx.cbuc.es

- Biblioteca Virtual M. de Cervantes www.cervantesvirtual.com/tesis/tesis_catalogo.shtml

- Dart Europe - http://www.dart-europe.eu

3 Bibliographical Review

Researchers from different scientific fields are working on the role of Twitter in politics; political scientists, sociologists, communication scholars, computer scientists, etc. Thus these authors come from different theoretical perspectives, use different methods and published in different venues. This leads to a rich but loosely interconnected body of evidence on the use of Twitter in politics. In order to achieve the objectives of the paper I followed Jane Webster guides for her paper on identifying the relevant literature on the topic of IS (Information Systems) (Webster & Watson, 2002). Data mining in Social Media,

likewise IS, is a new technological and interdisciplinary field straddling other disciplines so I found appropriated to follow her guides which can be summarized in:

Step 1 - The major contributions are likely to be in the leading journals, books and conferences. It makes sense, therefore, to start with them. Towards this end, I started a search in my sources of information using different combinations of the keywords: Twitter, predict, political, forecast, social media, survey, polls and review and I collected an initial core of articles. Then I checked the resulting studies for their correspondence with the topical focus of this review.

Step 2 – I went backward by reviewing the citations for the articles identified in step 1 to determine prior articles you should consider.

Step 3 - I went forward by using Google Scholar to identify articles, related with my topic, citing the key articles identified in the previous steps.

I thereafter studied and filtered this initial pool of articles in order to come up with a core of key documents plus a set of related documents which conform the final set that was included in my research. These papers are important because they are either the most quoted or the ones that reference the most significant works. The authors of these core articles are:

- Adam Bermingham (Bermingham & Smeaton, 2011)

- Sheng Yu (Yu & Kak, 2012)

- Daniel Gayo-Avello (Gayo-Avello, 2013)

- Nugroho Dwi Prasetyo Thesis (Prasetyo, 2014)

- Kalampokis Evangelos (Evangelos, Efthimios, & Konstantinos, 2014)

- Scott A. Golder (Golder & Macy, 2014)

- Andreas Jungherr (Jungherr, 2014)

- Daniel E. O'Leary (O'Leary, 2015)

- Tsakalidis (Tsakalidis, Papadopoulos, Cristea, & Kompatsiaris, 2015)

- David Vilares (Vilares, 2016)

All these literature, regardless of their objective, include an extensive and well documented concept-centric bibliographical review; e.i. concepts determine the organizing framework of each review and are a perfect entrance to any Social Science researcher interested in the presumed power of state-of-the-art methods and techniques for Twitter-based electoral prediction. From the literary review of this core articles', in the "Results" section I have detailed a summary of the algorithms used for data collection in Twitter plus a synthesize list of the main challenges and issues that have to be tackled in data collection for electoral prediction with Twitter.

To complement these approaches, I have taken an author-centered approach to the whole identified literature, not just the core articles'. This method may fail to synthesize the literature but the former list of core authors and its most relevant papers (which are already cited side by the author) are an excellent source of information and, to produce again a mere concept-centric review will be redundant. Besides, as stated in specific goals[3] I had to create a Delphi panel for variables B) and C) therefore I had to find out the key researchers and the different researcher profiles involved in this research. Because of this, an author-centered approach is more appropriated and complements the concept-centric approach that the other researchers did.

The snowball progression from the core authors resulted in 245 names and a set of 152 related articles[4]. There is no guarantee that this review covers all relevant studies on the topic, and papers in

3 - Can methods like data-mining, sentiment analysis or similar be use in Spain more successfully then offline polls?

- Are Spanish Social Scientist prepare to use those methods?

4 For the sake of clarity, the list of the most relevant articles is presented at the end of this paper in the Literature Review References section

languages different that English or Spanish have been excluded. Still, the combination of database, Google and snowball searches, covers a very significant part of the relevant literature.

4 Network and Key People

To obtain a clear picture of the patterns of scientific collaboration, key people, clusters and academic areas involved in these research area and, assess the bibliographical review that has been done, the researchers database has been formatted, imported to Gephi with the help of Sci2 and rendered as a co-authorship network with information about researchers, authoring and academic field. I expect the different measures of centrality plus some basic descriptive statistics to provided me with a better view of the field, highlight the most common researcher profiles and display the key researchers that have to bee invited to join the Delphi panel.

The parameters that I need from Gephi to determine the research landscape are:

Degree and Avg. Weighted Degree: edges and the proportion of existing edges count relative to maximum possible edges count. It highlights the most productive researches in a co-author-ship network.

Closeness Centrality: in a network, the distance of two nodes is the length of the shortest path between them. The farness of a node is the sum of its distances to all other nodes. It examines how well connected is a node (a researcher) with all other (Freeman, 1978) The lower the value, the most connected the node is, i.e. the parameter may display the most relevant authors.

Betweenness Centrality: the betweenness centrality quantitatively measures the control of a node on the communication between other nodes in the social networking. It measures the influence of the different nodes-researchers. Higher values will correspond to the gatekeepers, researchers that bridge between different clusters of researchers.

From the database I expect to have information about the nationalities and academic profiles and research. The network information mapped over the statistical information allows me to find out the most relevant Spanish researchers in this field and invite them to the Delphi panel.

5 Delphi Method

This technique was developed at the Rand Corporation in the early 1950s and is an efficient and effective group communication process that avoids many of the psychological distractions inherent to roundtable discussions. It was designed to systematically elicit judgments from experts in their selected area of expertise without those being influenced by the other participants in the panel (Ono & Wedemeyer, 1994)

The Delphi method was originally developed as a systematic, interactive forecasting method, however several studies have found that the most promising uses of a Delphi exercise are not in the realm of future prediction, but rather in value and panel analysis. The Delphi technique can be useful in gathering opinions from a large number of people in order to provide data for formulating organizational goals and discover areas of consensus. The Delphi technique is particularly well suited for the investigation of areas in which no real models exist and for which hard data is either insufficient or inconclusive (Aharony & Bronstein, 2014). The present study was conducted along this line of thought. This study does not try to ‘‘predict’’ the future of forecasting with data-mining, but rather attempted to formulate new goals and objectives that will help Social Scientist to reach an understanding on developing trends in the field. The panel in the current study included Spanish academic and professional experts. The academic sample was rendered out of the literary review (database plus network) and is a list of all the Spanish researchers that has published at least one article on using Twitter data to obtain knowledge about our society. The professional sample was one of convenience, and experts were professionals working in some Spanish companies or public organizations.

Results

The current section assesses the results and has two main purposes. The first one is to display and evaluate the results. The second purpose is to pipe the results from one section to the other one. E.g. the results from the academic review produced a list of researchers that have published a paper and a list of common features in those papers. The list of features is used to detect a general working algorithm, challenges, caveats and the weaknesses that should be tackled and that any Social Science researcher has to know. The list of researchers has been feed in Gephi to render a co-author ship network. Out of this network I have reviewed the researchers profile and selected the experts for the Delphi panel. Out of the list of challenges, caveats and weakness plus the information from the network I have created the survey for the Delphi panel.

Therefore this section is structured in a pipeline form: literary review results -> co-authorship network results -> Delphi panel results.

1 Assessments on the bibliographical reviews

a) Algorithms

"Although unstated in most of the studies, it is commonly implied that any method to predict Electoral results from Twitter data is an algorithm. Such algorithms are devised as a pipeline that starts with the collection of data, goes on processing that data, and finishes with a prediction that needs to be evaluated against the actual results of the elections. Needless to say, the algorithms can be parameterized to adapt to different scenarios, and predictions can be more or less detailed. Thus, there are a number of features defining any method to predict electoral results from Twitter" (Gayo-Avello, 2013). Almost all the concept-centric literary reviews in data mining for political prediction follow these assumptions, therefore I have synthesize the relevant ones.

Daniel Gayo resume these features as:

- **Step 1- Period and method of collection**: that is, the dates when tweets were collected and the parameterization used to collect them.
- **Step 2 - Data cleansing measures**:
  - Purity: i.e. to guarantee that only tweets from prospective voters are used to make the prediction.
  - De-biasing: i.e. to guarantee that any demographic bias in the Twitter user base is removed.
  - De-noising: i.e. to remove tweets not dealing with voter opinions (e.g., spam or disinformation) or even users not corresponding to actual prospective voters (e.g., spammers, robots, or propagandists)
  - Prediction method and its nature:
    - The method to infer voting intentions from tweets.
    - The nature of the inference: i.e. whether the method predicts individual votes or aggregated vote rates.
    - The nature of the prediction: i.e. whether the method predicts just a winner or vote rates for each candidate.
    - Granularity: i.e. the level at which the prediction is made (e.g., district, state, or national)
- **Step 3 - Performance evaluation**: i.e. the way in which the prediction is compared with the actual outcome of the election.

Nugroho Dwi Prasetyo ((Prasetyo, 2014) has added another main step which is implicit in Gayo approximation (Gayo-Avello, 2013);

- **Step 1 - Data collection**: Contains information such as selected API type, the number of tweets/user, the duration for collecting the data, and the keywords/hashtags. There are two methods on how to connect and collect tweets from Twitter; by searching tweets matching to the

keywords or by collecting all the tweets provided by Twitter through streaming API, or all the tweets in a specific language, or all the tweets in a specific location and then put all of them into the database. Period of data collection, keywords and type of election are other variables that have been considered.

- **Step 2 - Data filtering**: focus on cleaning the data such as deleting spam, non-political tweets, and removing non-potential voters, etc.
- **Step 3 - De-biasing of the data**. Addressing data bias is an important aspect in this process. Users of the social media do not represent the global population. Because of that, several research has tried to determine the demographic strata where the users belong to and weighting their tweets accordingly before the calculation process.
- **Step 4 - Prediction calculation**. Based on the features used by the authors, we can divide the calculation methods into two main categorizes, parameter count and sentiment analysis. In parameter count, counting tweets is the most common feature used by researchers followed by counting re-tweet, user, and interaction between candidate and potential voter. The second category is applying sentiment detection in each tweet to classify positive, negative and/or neutral tweets. This could be performed by using several approaches such as lexicon-based, supervised machine learning, and crowd-sourcing.

Kalampokis Evangelos (Evangelos et al., 2014) has created a similar framework. His proposal comprises two discrete phases and each of these phases can be further divided into a sequence of stages and each stage into a number of steps:

Step 1 - Data Conditioning Phase. Refers to the transformation of noisy raw Social Media (SM) data into high quality data that is structured based on some predictor variables.

- Collection and Filtering of Raw Data. This stage deals with both raw SM data collection from various sources and filtering of data in order to determine those relevant. The steps composing this stage are the following: Determination of time window. Identification of location. Identification of user profile characteristics. Selection of search terms.
- Computation of Predictor Variables. This stage deals with analysis of the raw data resulting from the previous stage in order to compute the values of predictor variables. In this stage, only variables related to SM are considered despite the fact that more variables (e.g. product price) can be finally employed in the predictive analysis stage. The steps composing this stage are the following: Selection of predictor variables. Measurement of predictor variables. Computation of predictor variables.

- **Step 2 - Predictive Analysis Phase**. Refers to the creation and evaluation of a predictive model that enables estimating outcome from a new set of observations.
  - Creation of Predictive Model. In this stage the actual model is created based on statistical or data mining methods. The steps composing this stage are the following: Selection of predictive method. Selection and use of non-SM predictor variables. Identification of data for evaluation of prediction.
  - Evaluation of the Predictive Performance. In this stage prediction accuracy is evaluated against the actual outcome. The steps composing this stage are the following: Selection of the evaluation method. Specification of the prediction baseline.

Although so far there is no standard procedure for any of the steps involved. Procedures for data collection and sampling vary widely across disciplines, and often adapt differently to the peculiarities of the platforms from where the data are gathered or when third party solutions are used to gather the data. As a result, it is difficult to integrate research outputs and this hampers the ability to replicate findings and engage in cumulative research and theory building (Gonzalez-Bailon, Wang, Rivero, Borge-Holthoefer,

& Moreno, 2012) But those are the features that the authors have identified and assessed from the different papers reviewed and can be used by Social Science researchers in order to decide or compare whether or not a planned methodology is feasible for political prediction.

Andreas Jungherr (Jungherr, 2014) didn't center his review only in political prediction and has a more descriptive approach to the different papers on the subject of Twitter and politics. His review includes a discussion of the theories, research designs, methods of data collection and data selection that were most common in the literature included in the review, plus a short synopses of all the included studies. It's an excellent source of information that combined with his “Tutorial for Using Twitter Data in the Social Sciences” (Jurgens & Jungherr, 2016) provides Social Science researchers with and excellent toolkit to begin with political prediction using data from Twitter.

b) Data Collection on Twitter; issues and challenges

After reviewing the literature, the main challenges, caveats and the weaknesses that should be tackled in data collection in Twitter for political forecast within the future are.

1. It's too dependent on arbitrary decisions such as the parties or candidates to be considered, or the selection of a period to collect the data. (Jungherr, Jurgens, & Schoen, 2011), (Gayo-Avello et al., 2011)
2. We should not disregard the hypothesis that positive results could have been due to chance or, even, to unintentional data dredging due to post hoc analysis. (Gayo-Avello, 2012a)
3. Sentiment analysis is applied as a black box and with naïveté (Gayo-Avello, 2012a). Sometimes commonly used methods are slightly better than random classifiers. (Cameron & Barrett, 2013)
4. Methods have to be able to take humor and sarcasm into consideration most of them fail to catch the subtleties of political discourse. (Gayo-Avello, 2013; Liu, 2012; Wlodarczak, Soar, & Ally, 2015)

5. All of the tweets are assumed to be trustworthy which it's not the case. Spam, misleading propaganda and astroturfing should be detected and filtered out or the method should be tolerant to that noise. (Gayo-Avello, 2013)
6. Demographics bias is ignored even when it is well known that Social Media data is not a random or representative sample of the population (Duch, 2016; Mislove et al., 2011; Morstatter, Pfeffer, Liu, & Carley, 2013) besides up until now, there is only a limited number of studies that have large enough samples compared with the population of the area under study (Jungherr, 2014)
7. Gender, ideology or party identification of Twitter users are some of the many latent variables that cannot be observed directly (Barbera & Rivero, 2014) but have to be assessed because the inequality can be really deep (Barbera & Rivero, 2014; Barbera, 2016; Diaz, Gamon, Hofman, Kiciman, & Rothschild, 2016; Mislove et al., 2011)
8. Self-selection bias is simply ignored. People tweet on a voluntary basis and, therefore, data are produced by those politically active and/or with extreme values in the ideological scale. (Barbera & Rivero, 2014) (Gayo-Avello et al., 2011) (Conover et al., 2011; Diaz et al., 2016)
9. The set of messages retrieved through the APIs is not random sample of all Twitter activity. There is no systematic account of the bias unless you have access to the full stream of activity (or firehose which is out of reach for most research organizations (Gonzalez-Bailon et al., 2012) furthermore, there are distinct network topologies depending the keywords being used or the mechanisms (mentions versus retweets) analyzed (Conover et al., 2011)
10. Tweet based prediction is not applicable in the same way in all parts of a geographical area or even inside a State. Employing the same prediction model, the MAE can vary from 0.05% to 25.01% in the provinces where the number of users are less than 5 thousand; and city size or rural area versus urban area can be an important source of inequality in the user representation (Prasetyo, 2014) (Gaurav, Srivastava, Kumar, & Miller, 2013) (Vilares, 2016) (Barbera & Rivero, 2014) and this is something that it does not seem to go towards a more balanced distribution "Twitter is more geographically biased now [21013] than it was in 2008" (Gordon,

2013). Although Pablo Barbera found no significative differences in Spain between twitter users from spanish big cities and the other ones (Barbera & Rivero, 2011)

11. In many occasions, election results were determined by the choice of swing voters. While it is very easy to detect their share in offline/interview-based polling, it is hard to detect them based only on tweets. Accurate socio-demographic information is crucial (Murthy, 2015)
12. If online and social media data are to treated as surveys, they must be treated as imperfect surveys. “Traditional surveys follow a rigorous procedure, asking the same question (...) to repeated cross sections of a random sample of a representative group of people. A search and social “survey,” however, is essentially polling a varying, non-random sample of voluntary participants who selectively respond to questions of their choice” (Diaz et al., 2016)
13. The language may be a problem. Many powerful language processing technique are English-based. Languages like English, Spanish or French are so well established all around Internet that sometimes is difficult to filter out keyboards related with political subjects of one country, from all the other ones. This makes the process of language model building become more difficult, and lowering the classification accuracy. (Barbera, 2016; Prasetyo, 2014)
14. “For the field to advance beyond the state of monadic case studies (…) it is necessary to use the available evidence and establish a common ground of usage patterns and phenomena that are to be expected for various aspects of the political use of Twitter” (Jungherr, 2014)

2 Assessments on the coauthor-ship network and database.

Database

Political trends prediction is an academic field clearly related to Political Science or Sociology, but the database of researchers created after the literary review

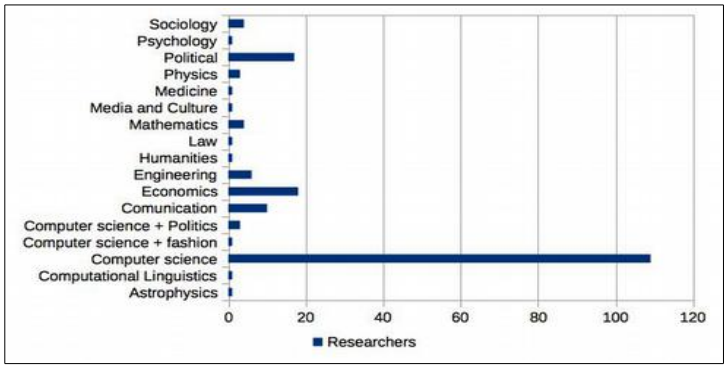


Image - 1

shows that 90% of research being done in this field is being conducted by researchers related to Computer Science (Image 1)

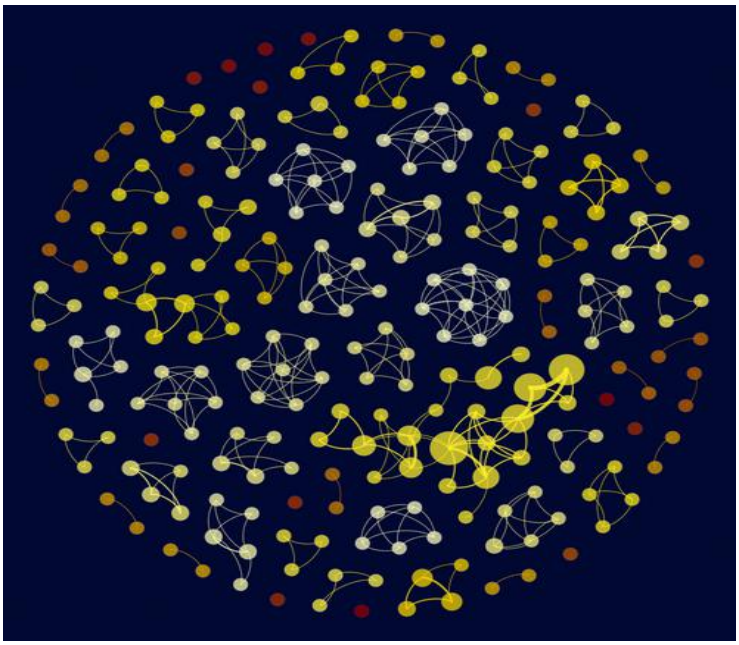


Image-2

Network

The consequences of this are clearly seen in this image of the co-authorship network (Image 2) This image displays all of the researchers that have published at least a paper related with using Twitter data to forecast electoral prediction or political trends. Each researcher is a node/vertex. The vertex size is proportional to the productivity of the authors, while the thickness of the lines indicates the strength of the connection between two authors and the edges show the co-authorship relations. The map shows many researchers working alone or in small groups, without any relation to other researchers and with very few papers published. It is a map full of small, absolutely unconnected, islands. There is only a small area where we can observe a larger structure that looks more like a research network with nodes of different sizes and edges of different intensity.

The analysis of Image-1 and Image-2 seems to indicate that, since most of the researchers are not from the field of Social Sciences, the research that they do in this area is merely punctual.

Degree and Avg. Weighted Degree(AWD)

Degree counts de edges and AWD the proportion of existing edges count relative to maximum possible edges count. It highlights the most productive researches in a co-author-ship network. Avg. Weighted Degree display how often researchers collaborate with each other.

Researcher	Number of authored works	Academic field
Gayo-avello, Daniel	9	Computer Science
Jungherr, Andreas	8	Political Science
Schoen, Harald	7	Political Science
Jurgens, Pascal	7	Comunication
Mustafaraj, Eni	5	Computer Science
Barbera, Pablo	5	Political Science
Mendoza, Marcelo	4	Computer Science
Poblete, Barbara	4	Computer Science
Bravo-marquez, Felipe	4	Computer Science
Bollen, Johan	3	Psychology
Takis Metaxas, Panagiotis	3	Computer Science
Ceron, Andrea	3	Political Science
Curini, Luigi	3	Political Science
Burnap, Pete	2	Computer Science

Researcher	Degree	Academic field
Gayo-avello, Daniel	10	Computer Science
Schoen, Harald	8	Political Science
Mustafaraj, Eni	7	Computer Science
Bu, Jiajun	7	Computer Science
Chen, Chun	7	Computer Science
Guan, Ziyu	7	Computer Science
He, Xiaofei	7	Computer Science
Li, Yang	7	Computer Science
Sun, Huan	7	Computer Science
Tan, Shulong	7	Computer Science
Yan, Xifeng	7	Computer Science
Mendoza, Marcelo	6	Computer Science
Poblete, Barbara	6	Computer Science
Caldarelli, Guido	6	Physics

Researcher	Weighted Degree	Academic field
Mustafaraj, Eni	22	Computer Science
Gayo-avello, Daniel	18	Computer Science
Chen, Chun	15	Computer Science
Schoen, Harald	14	Political Science
Bu, Jiajun	14	Computer Science
He, Xiaofei	10	Computer Science
Li, Yang	10	Computer Science
Yan, Xifeng	10	Computer Science
Chessa, Alessandro	10	Physics
Pompa, Gabriele	10	
Sun, Huan	9	Computer Science
Tan, Shulong	7	Computer Science
Caldarelli, Guido	7	Physics
Pammolli, Fabio	7	Economics

Closeness Centrality measures the length of the shortest path between two nodes. The farness of a node is the sum of its distances to all other nodes. It examines how well connected is a node (a researcher) with all other. This parameter may display the most relevant authors, but because it's an average value, The values of the nodes in isolated groups are equal to those of very large groups of connected nodes. That is, in a network of only two connected nodes A and B their Closeness will be 1 whereas in a network of 100 nodes, a node C with a grade 20 will have a value of 5 so it might seem that A and B are nodes better connected then C which is not the case. We saw in Image - 2, that there are many small groups and solitary researchers, therefore Closeness Centrality will not be useful because it will produce mixed values. So instead I have used Betweenness Centrality which quantitatively measures the control of a node on the communication between other nodes in the social networking.

Betweenness Centrality

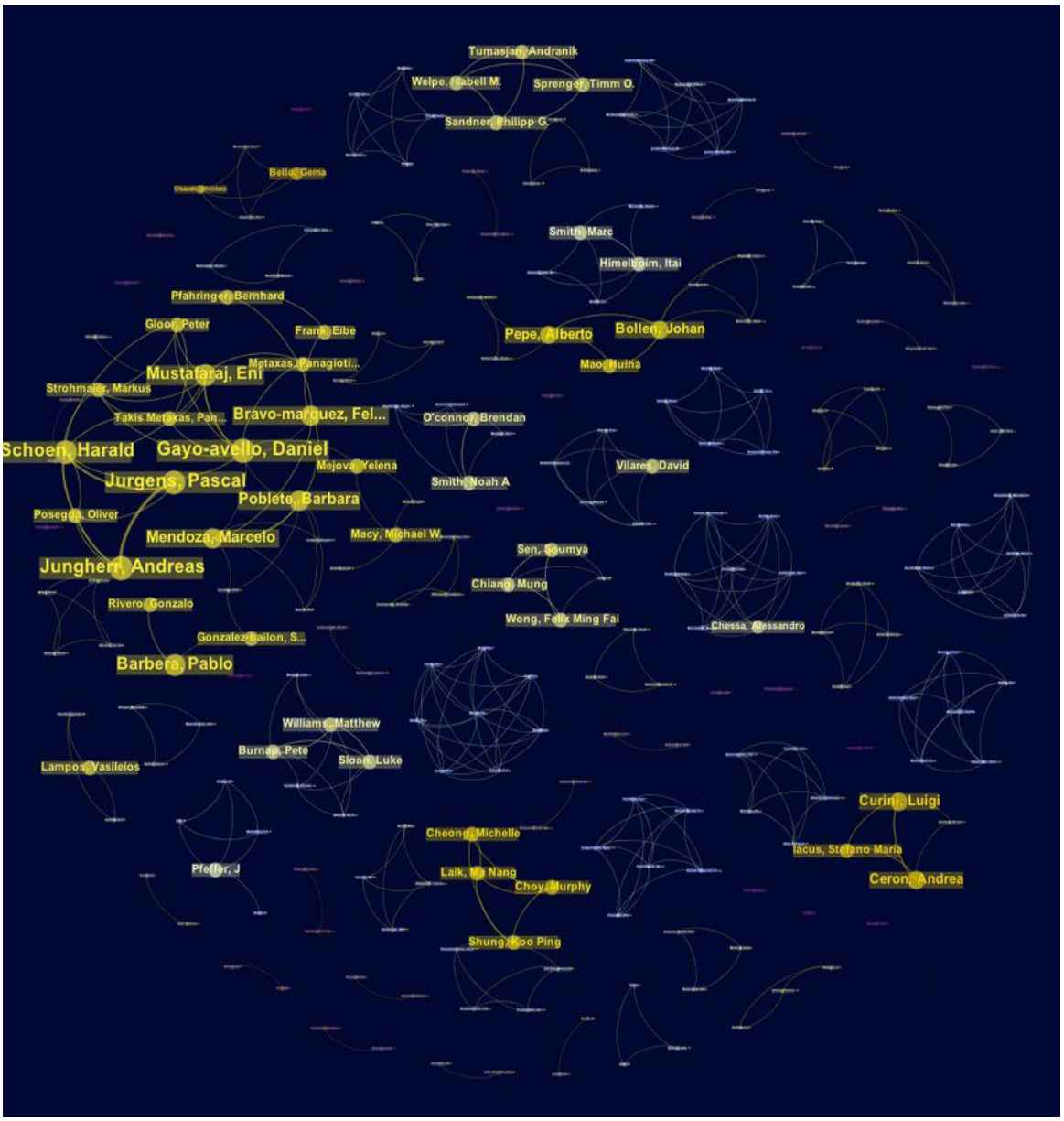
It measures the influence of the different nodes-researchers. Higher values will correspond to the gatekeepers, researchers that bridge between different clusters of researchers.

Researcher	Betweenness Centrality	Academic field
Gayo-avello, Daniel	0,007	Computer Science
Schoen, Harald	0,002	Political Science
Mejova, Yelena	0,002	Computer Science
Castillo, Carlos	0,002	Computer Science
Mendoza, Marcelo	0,002	Computer Science
Poblete, Barbara	0,002	Computer Science
Bravo-marquez, Felipe	0,002	Computer Science
Gonzalez-bailon, Sandra	0,002	Sociology
Mustafaraj, Eni	0,001	Computer Science
Macy, Michael W.	0,001	Arts
Barbera, Pablo	0,001	Political Science
Bollen, Johan	0,000	Psychology
Pepe, Alberto	0,000	Astrophysic
Pfeffer, J	0,000	Computer Science

Now, mapping the relevant researchers over the nodes of the network we render this image-landscape of the research in data mining Twitter for political prediction Image-3

Image-3 shows that most of the few Social Scientists: Jurgens, P.; Schoen, H; Barbera, P.; Rivero, Gonzalo; Jungherr, A. and Gonzalez-Bailon,S.; are part of this big cluster and there are some Computer Scientists which sort of act as gatekeepers; Gayo-Avello, D.; Mejova, Y.; Poblete, B.; Bravo-Marquez, F.; Mustafaraj, E.; Mendoza, M.; Castillo, C. bridging between the Social Scientists. So apparently is possible to create a collaborative network of research between Social Science and Computer Science and this one is more prolific and dynamic then the small, homogeneous and monadic research teams.

Image-3



3 Assessments on the Delphi panel

This part examined the experts views concerning three main themes. The first theme includes a number of general issues regarding the datasets and its bias. The second one focus on the use and validity of twitter datasets for electoral prediction in Spain and the third theme focus on what knowledge Social Science students may need and whether or not, the Spanish students have this knowledge.

In Spain there are about 49 academic experts on this area, 39 have been invited to participate (full list in Appendix 0) 22 responded and 18 accepted to participate in the survey, resulting in a 46 % participation rate, a very high one and an appropriate group size for a Delphi questionnaire according to the academic literature (Huertas, P. L., Moro, A. I., & López, 2005; Huertas & Moro, 2006) .

An online survey (see Appendix 1), specifically designed and constructed for the present study, contained a 21-statement online questionnaire intending to reflect the main issues and trends found in the literature reviewed for this study[5]. The statements were rated on a six-point Likert-type scale to avoid neutral positions: (0) very improbable/very undesirable or negative; (1) improbable or unlikely /undesirable; (2) moderately improbable/ moderately unlikely (3) slightly probable/moderately desirable; (4) probable/desirable; and (5) very probable/very desirable or positive. An Average point for each variable was calculated to determine de tendency in this statement; 0.0 min. 2.5 neutral 5.0 max.

All the figures for the Delphi Panel are on Appendix - 2. Due to a tight deadline, I didn't have time to follow Huertas advice (Huertas, P. L., Moro, A. I., & López, 2005) about making some prior test in order to ensure the validity and usefulness of all the statements in the questionnaire and it was a mistake because some questions were misunderstood. The desirability response value in some statements

5 For most of the questions, consensus was reached in the first round. Planning for a second one was complicated and I stopped it because the Delphi panel was coincidental with two major elections (Brexit and Spanish Elections) and the end of the academic year.

was confused because some experts instead of rating the variable (b) “what they would like to see happen (desirability)” rated the desirability of the variable (a) “what they think will probably happen (probability)”. For instance, one statement says: “In Spain the knowledge in Statistics and Computer Science that social science students have it's correct” Some experts rated variable (a) as unlikely (1) and variable (b) as very positive (5) because they thought that is was positive for the students to have this knowledge. But others experts rated variable (a) as unlikely (1) and variable (b) as negative (0) because they considered that was very negative that the students didn't have this knowledge. So the rating was not for variable (b) but for the fact that variable (a) was very low. Both express the same; “students must have this knowledge”. But it create confusion in the rating. For this reason the ratings for variable (b) in the following statements are considered no valid: T-1 Statement 4, T-2 Statement 5, T-3 Statements 5 & 6.

Theme 1: Datasets and bias

The literary review have highlighted that an accurate prediction can only come through correctly identifying likely voters and getting an un-biased representative sample of them in order to be able to stratify them and weight ideological opinion estimates (Barbera & Rivero, 2014; Gayo-Avello et al., 2011; Gayo-Avello, 2012b; Gonzalez-Bailon et al., 2012; Jungherr et al., 2011). Because this is a key subject, the firsts three statements were precisely about “bias” and asked the experts whether they thought that bias will disappear in a short period of time, (roughly five years).

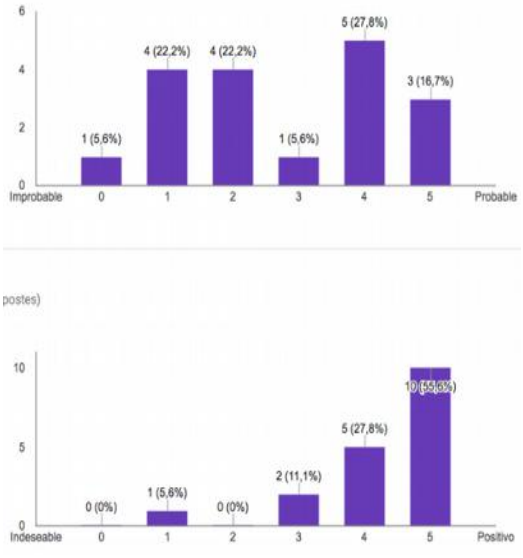


Image-4

Statement-1 (Image-4)

“In less than 5 years it will be possible to devise sorting algorithms and sentiment analysis techniques for text mining

which will be able to understand the subtleties, twists and idiosyncrasies of the language and detect spam, repetitions, trolls, irony, astroturfing, etc."

Participants viewed these as highly desirable (average point 4.28) but they had divided views about its probability (average point 2.78) with a broad range of ratings and 44.4 % considered it unlikely or moderately unlikely. One expert states that still in the most successful cases, it will never be possible to completely overturn this bias because even us, as human beings, have problems understanding irony, sentences without context or figurative language. Which is consistent with the academic literature. P. Wlodarczak says (Wlodarczak et al., 2015) "Languages are ambiguous and humor and innuendos cannot easily be analyzed using text mining techniques" and Liu (Liu, 2012) could detect sarcasm in only 56% of the cases.

But half of the experts where much more optimistic. The reason of this differences was in where the experts focus the advances. If an expert was more interested in humor, irony, sarcasm etc detection, he/she was more pessimistic. If the expert was more interested in spam, astroturfing and troll detection he/she was more optimistic. Therefore, for a future research it will be better to first identify the common characteristics of different text mining goals and subdivide text mining in different similar areas.

Statement-2 (image-5)

"In less than 5 years, thanks to different techniques or the use of external datasets, it will be possible to eliminate demographic biases and obtain accurate socio-demographic profiles of Twitter users".

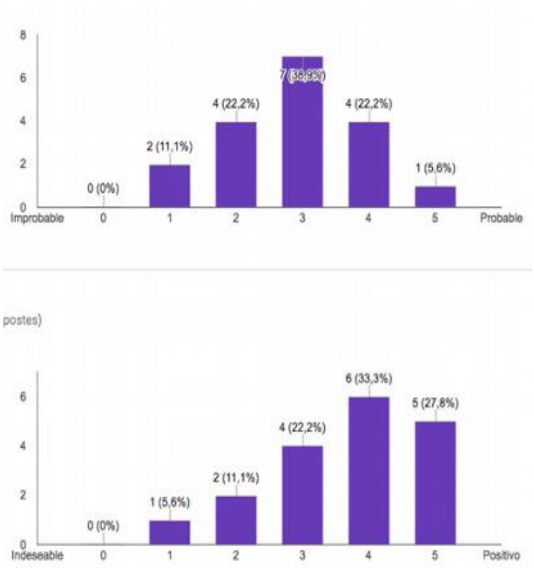


Image-5

Researchers were also dubious about the possibility of obtaining more background information about each individual user, although in this statement there was more consensus and they consider it slightly

probable (avg. Point 2.89) On the other hand, the experts saw it least desirable (avg. Point 3.67). Probably the reason of this slightly more negative perception is due to some concerns, not related with academic research, but which may affects this research, as one expert pointed:

"*Las cuestiones de privacidad impedirán que se llegue a obtener toda esa información de todos esos usuarios. Hay muchas cuestiones legales y éticas que deben ser tenidas en cuenta*". Same concerns are arise by Golder and Macy: "These new sources of data raise challenging procedural, legal, and ethical questions about how to protect individual privacy".

In addition, one expert disagreed with the group consensus, stating: "Conforme pasen más años, los usuarios de twitter irán cubriendo todo el espectro de edades, pero no por técnicas de extracción de información, sino por pura ley natural."

Yet, these results contrast the findings of (Golder & Macy, 2014) who claim that "rapid progress is being made to address these limitations". In fact, most of the academic literature seems to agree with Golder findings and perceive that a huge progress is being made. Fernando Diaz (Diaz et al., 2016) comments "the processes for matching demographics are improving rapidly as users provide more profile information, more datasets are linked between individuals, and methodology for analytics improves". Some researchers had extracted occupation information from Twitter user profiles and conducted text analysis to categorize users into occupational classes (Lampos, Preotiuc-Pietro, & Cohn, 2013; Sloan, Morgan, Burnap, & Williams, 2015) and Jernigan and Mistree (Jernigan & Mistree, 2009) and Schwartz (Schwartz et al., 2013) showed how social media content from Facebook ot Twitter can be used to infer a wide range of user attributes, including age, gender, sexual preference, and political party affiliation. Schwartz et al., 2013)

Related with geo-location (Compton, Jurgens, & Allen, 2015) showed how label-propagation algorithms can be adapted to potentially geotag most of Twitter users within a few kilometers. (Duch, 2016) also states that "Accurate estimates of the most relevant sociodemographic characteristics of Twitter users – those that are often used to recover the representativeness of survey respondents – can be

accurately predicted from the text of their tweets and from who they decide to follow" and Pablo Barbera in one of his last papers (Barbera, 2016) says "accurate estimates of the most relevant sociodemographic characteristics of Twitter users – those that are often used to recover the representativeness of survey respondents – can be accurately predicted from the text of their tweets and from who they decide to follow". So, it looks like spanish experts are more cautious about future advances.

Statement-3 (image-6)

The obvious solution to avoid those bias could be to obtain unbiased information straight from de SNS, thus Statement 3 says:

"In less than five years the population samples that Social Media companies will provide will have no unknown biases".

As grap t1-S3 shows, the experts were extremely skeptical about this.

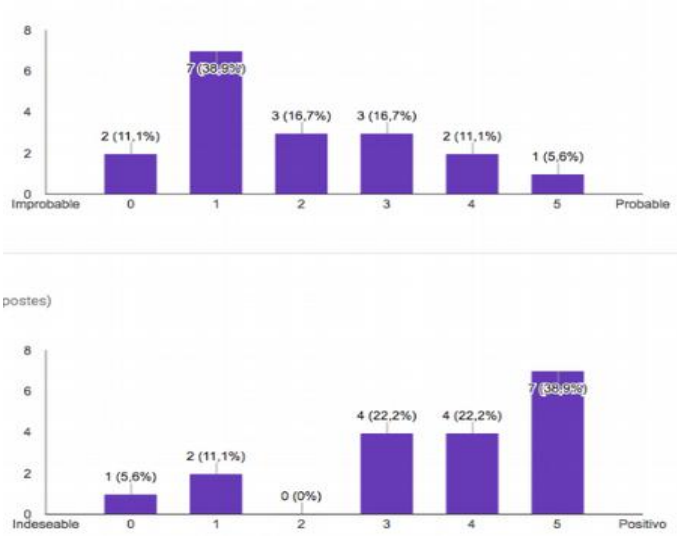


Image-6

Regardless of the likeliness to obtain the sociodemographic characteristics of Twitter users in the near future, all experts agree that nowadays there is still a lot of research to be done in order to be able to surpass all the bias before mentioned. But research in data mining requires access to good, rich and structured os semi-structured datasets with data that can be used as benchmarks. So the last four statements in this first part are about the possibility of this datasets and about the researchers accessing them.

Statements - 4 - 5

Statement 4: "In less than 5 years it will be possible to access for free all data from users of SNS"

50% of experts found it very unlikely and the avg. point stays at 1.0 so there is a very negative view of SNS charging for researchers accessing user data but as one expert says: "The data from users of

Social Networks have a great value for the companies that manage these networks and they will not give it away for free" another expert states that to sell user profiles will be a big business pretty soon.

Statement 5: "In less than 5 years the costs to access data from users of SNS will increase"

Half of the experts consider it quite probable with an avg. point of 2.89 and almost all of them (avg. point 0.94) found it very undesirable. There is a consensus about the fact that it will be increasingly difficult and more expensive to access all the data that SNS have and this goes linked with Statement five. So as soon as the SNSs find a way to transform this data in assets they will manage it as another asset with an economic value.

Statements - 7 (image-7)

"Soon partnerships or joint venture of Social Media Sites, Telecoms and Financial corporation will create datasets very precise and structured not open to researchers"

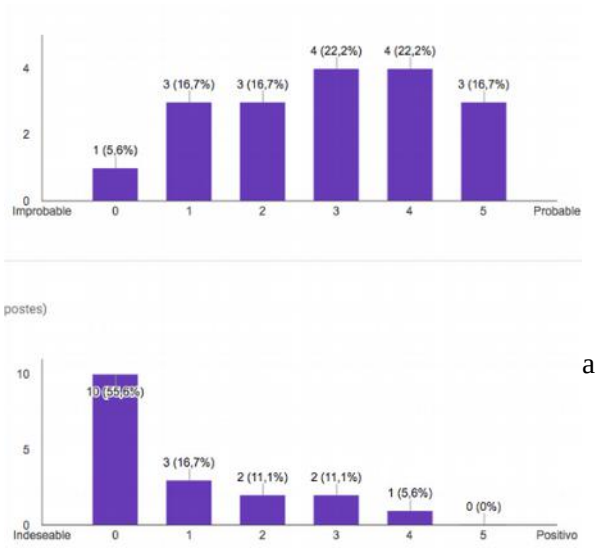


Image-7

There is a consensus (avg. point 4.11) about the probability that this private datasets will emerge and a very negative view 55% about it. So, researchers, in order to do research, need big datasets with structured or semistructured information and SNS and corporations are creating this, datasets but this datasets will be increasingly difficult to access because its information is private and considered to be a commercial asset. So could it be possible to access this datasets without compromising its commercial value? Next Statement focused on this problem.

Statement-8 (Image-8)

"The large public or private datasets can be anonymized and provided to researchers to enhance research and obtain knowledge"

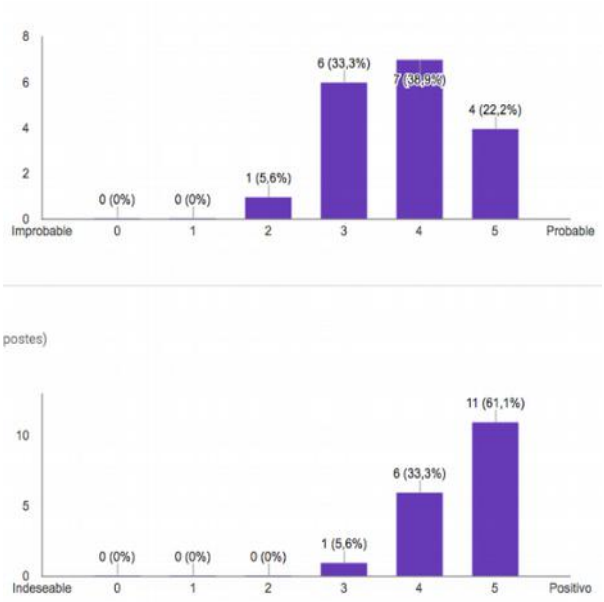


Image-8

The views expressed by participants regarding Statement 8 are quite interesting. There is a general agreement (avg. point 3.78) that big datasets could be anonymized and handed to researchers for academic research and almost everybody thinks that this will be either moderately desirable or very desirable (more then 90%). But there is a growing body of research showing that anonymizing or encrypting data is not sufficient for protecting privacy, as this can sometimes be reverse-engineered. So probably the owners of the datasets will be extremely reluctant to offer their datasets for public research. Maybe, as one expert wrote, the solution will come from national public policies either pushing forward to open access to this datasets or protecting its ownership:

"It would be desirable, but it depends on the political interest in each country to promote scientific research on these issues. There can be large differences between one country and another."

Statement-6 (Image-9)

"In less than 5 years, commercial interest in data mining and "machine learning", in the papers will cause an increased use of "Black boxes" that hide or do not itemize parts of the research".

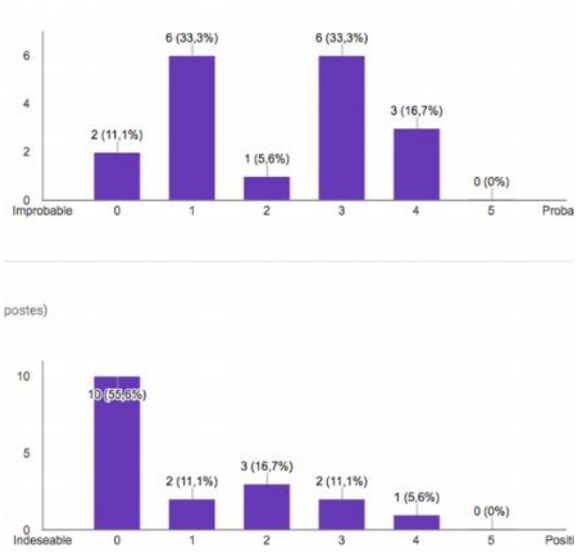


Image-9

The ratings show that there is indeed a concern about algorithms, or parts of the research, becoming sort of "black boxes" due to commercial interests. Almost half of the participants assume that the tendency toward "black boxes" is probable; and more then 82 % perceive this tendency as negative. But one expert wrote:

"*Una cosa es la explotación commercial (las empresas que se dediquen a estos ocultarán toda la información que puedan) y otra la investigación científica, que se basa en la reproducibilidad de resultados.*"

And another expert stated:

"Prestigious conferences will always require new methods to be published"

So there is concern but also hope.

Theme 2: About Spanish electoral predictions

Statement-1 (Image-10)

The first statement in this second theme is about the easiness of replicating research done in other countries to the Spanish context. Almost half of the participants assume that is probable (avg. point 3.06) and easy to do it; however more then 90 % perceive this as very desirable (avg. point 4.41). In other words, apart of having a similar political system which obviously may help, it looks like they would like to have more research works and papers to confirm the easiness or not of this adaptation.

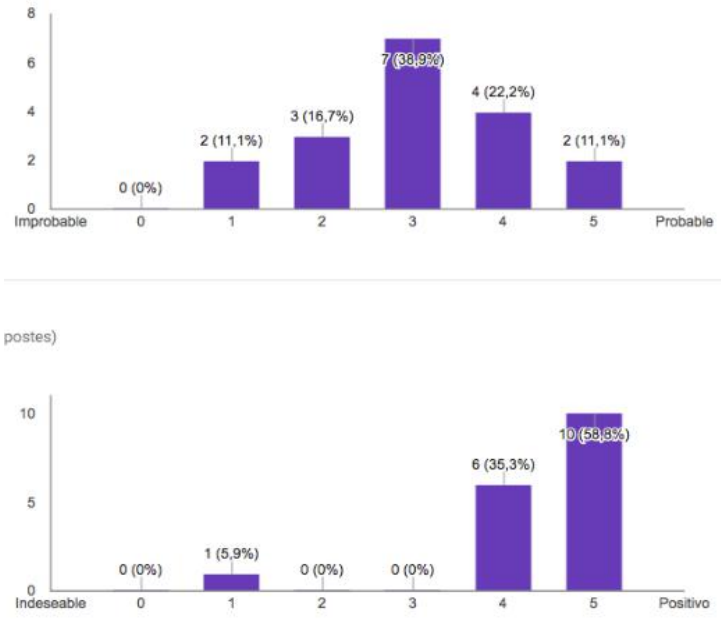


Image-10

Statement - 2 (Image-11)

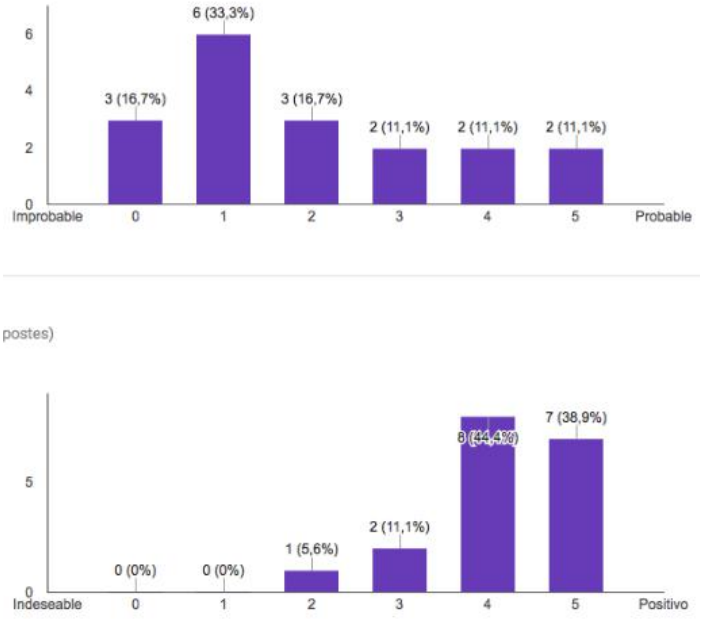


Image-11

''With the information available on the Internet soon it will be possible to make predictions on the outcome of all parties (regardless of their size or representativeness) with a MAE (Mean Absolute Error) equal to or lower than the polls''

This Statement evaluates the possibility that the size and representativeness (local parties versus national parties) of political parties may enhance or lower the accuracy of electoral prediction. The majority of experts (more then 66%) believe that it is unlikely, while almost all of them believed it is desirable (90 %). These findings echo the academic literature consensus which establishes that its desirable and it may be possible but currently Twitter data does not provide enough information to predict with accuracy the results of all the parties regardless of its size and representativeness.

Statement - 3 and 4

3 -"With the data available on Internet it's possible to make predictions about the great national parliamentary parties, with a MAE equal to or lower than the polls" (Image-12)

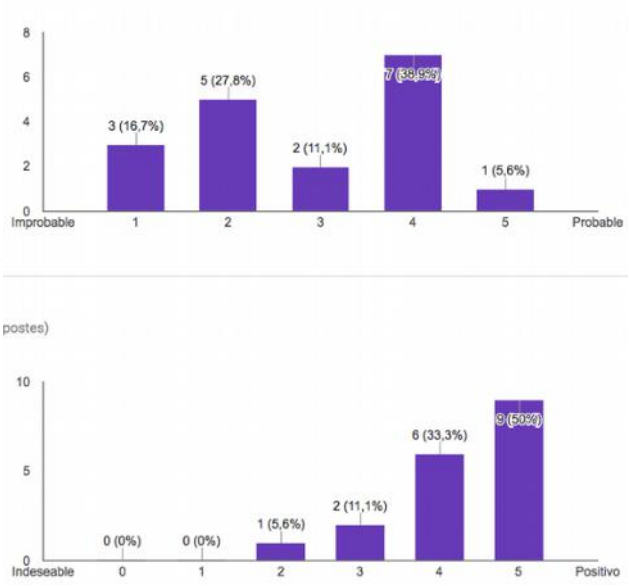


Image-12

4- “With the data available on Internet it's possible to make predictions of major trends or referendums involving many people with a MAE equal to or lower than traditional polls.” (Image-13)

Regarding this Statements all the experts found it very positive, and although the reality for the next years it's still not clear, (St-3 avg. point 2.89 and St-4 avg. point 3.00) there is a clear shift to the probable position. One expert who believes improbable Statements 2 but probable statement 4 explains that:

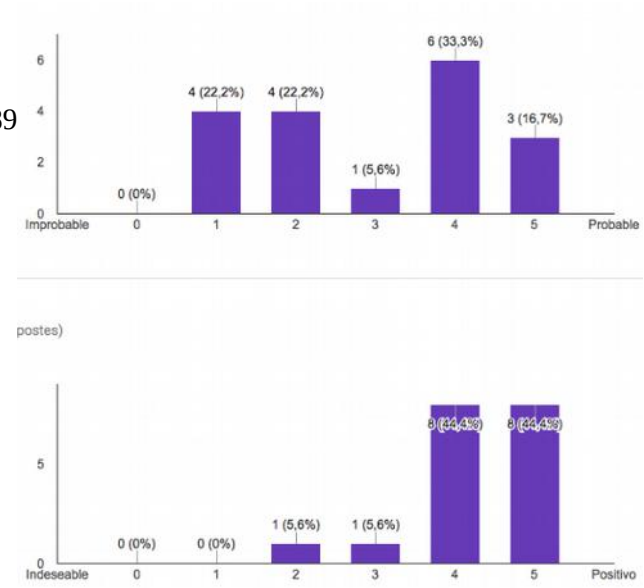


Image-13

“*It will be possible to poll with high reliability aspects of public opinion, in particular the opinion of the population on political issues, but the voting intentions of a Twitter user is still way to complex. The results obtained so far are not conclusive and in many cases even contradictory*”

Which is, with slight differences, a good sum up of all the experts opinion. When it comes to big trends detection, highly polarized opinions o even in major bi-polar referendums chances are that the accuracy of Twitter prediction will improve a lot. But to translate this to a particular vote and a precise electoral prediction for all the parties running in a national election is still a bit far away and it requires lots of work.

Statement - 5 (Image-14)

“In Spain there are private-public institutions with datasets that can be used to perform much more accurate analysis of political trends that the ones available to researchers”

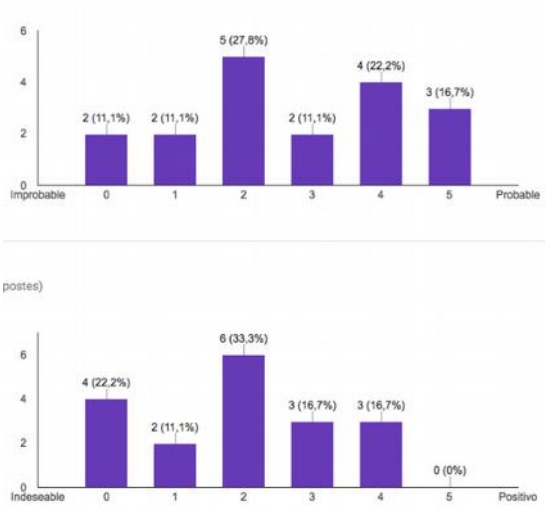


Image-14

Expert answers reflect an ambiguous attitude towards datasets and its potential use. While there is a perception that some institutions and corporations may own big datasets with good structured information as we saw it in Statement 7 Theme 1, only 50 % of participants believe that the owners of the datasets will be able to use them as a tool for political prediction.

Statements - 6 (Image-15)

"Politicians will create institutions to manage or limit the knowledge about social trends that can be obtained through data mining online datasets".

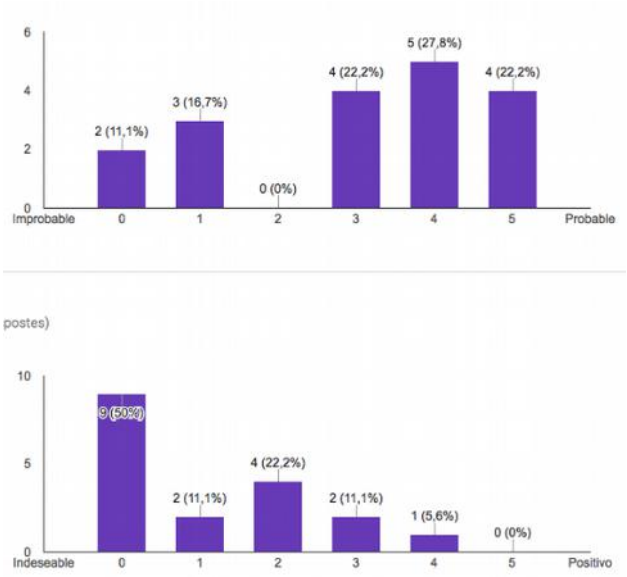


Image-15

The analysis of Statements 6 is intriguing as well as it suggests that, on one hand, experts assume that the tendency toward the creation of public institutions that will try to control/manage the potential knowledge that this datasets offer, will increase in the near future (avg. point 3.06) but, on the other hand, they just stated in the other Statements that: probably in a short period of time it will not be feasible to obtain precise information about electoral prediction, not even for the owners of good structured datasets. Furthermore the experts view the development of this institutions undesirable or very negative (avg. point 1.1) and half of them considering it very negative. So the question is, why are the experts so afraid about the political institutions creating institutions to control or manage something that so far it doesn't exist?.

Theme 3: Spanish education

Sandra Gonzalez (Gonzalez-Bailon, 2013) expresses perfectly the sense of this third theme, so I just quote her words: "The fact is that the volume of information does not reduce the role of human

interpretation, or the biases that we introduce in choosing the level, or resolution, for the analyses. One can conceive of scientific research as the institutionalization of the controls that put a bridle on those biases, or at least help us identify them. But for this, we still need theories and models, and cumulative research that works on the basis of improvements and readjustments. In other words, Big Data will not bring about the end of theory; quite the contrary. And social science has a crucial role to play in the discovery of the biases that are intrinsic to digital data, as well as in the construction of convincing stories about what those data reveal". Indeed, looks like Social Science has a crucial role to play in the Big Data world but, are Spanish Social Science researchers and students prepared?.

Statement - 1 (Image-16)

"The need to know programming tools such as R or Python or algorithms such as SVM, Random Forest, Naive Bayes, etc will increase soon''.

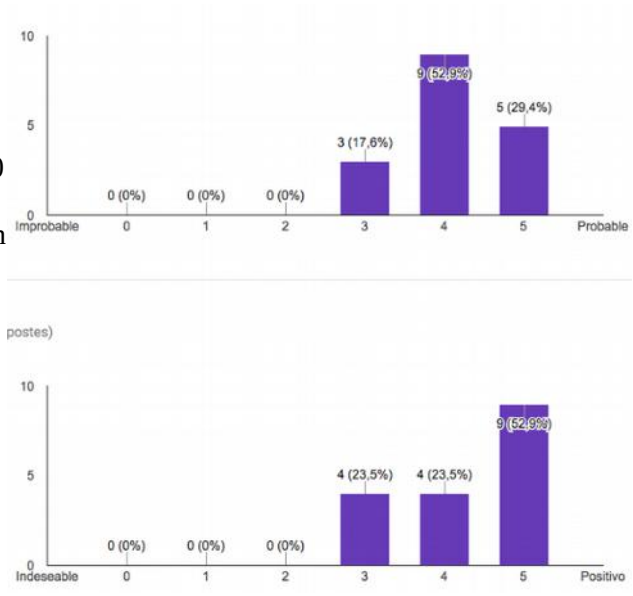


Image-16

The experts agree on the need to know programing tools and algorithms. None of them (100 %) found this trend unlikely and (100 %) found it -in different degrees- positive. These numbers accord with the already classical quote of Duncan Watts (Watts, 2011 - p.266)

"[J]ust as the invention of the telescope revolutionized the study of the heavens, so too by rendering the unmeasurable measurable, the technological revolution in mobile, Web, and Internet communications has the potential to revolutionize our understanding of ourselves and how we interact ... . [T]hree hundred years after Alexander Pope argued that the proper study of mankind should lie not in the heavens but in ourselves, we have finally found our telescope. Let the revolution begin".

Or, as Yelena Mejova expresses it (Mejova, Weber, & Macy, 2015b), without the automated data management and coding tools developed by computer scientists, the analysis of massive unstructured data will remain beyond the reach of most social scientists, leaving the field to disciplines that are much better at building powerful telescopes than at knowing where to point them.

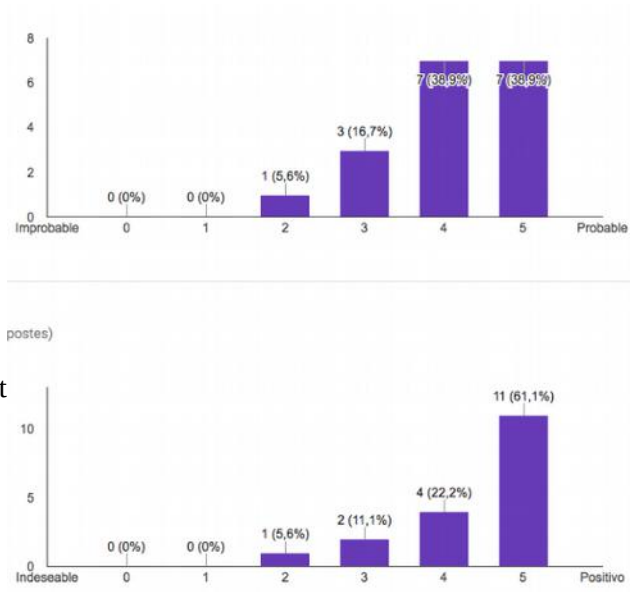


Image-17

Statement - 3 (Image-17)

"Soon, the power and possibilities of data mining tools and "machine learning" will be so great that new specialties will be created in sociology, humanities or political science focusing on managing the digital trail that people leave in Internet."

There is also a general agreement about it. More then 90 of the experts think that sooner or later it will probably happen and found it very positive. An expert said that there are already this departments, and indeed in some universities they already offer Digital Methods, Computational Social Science, or similar subjects, although is not common in Spain.

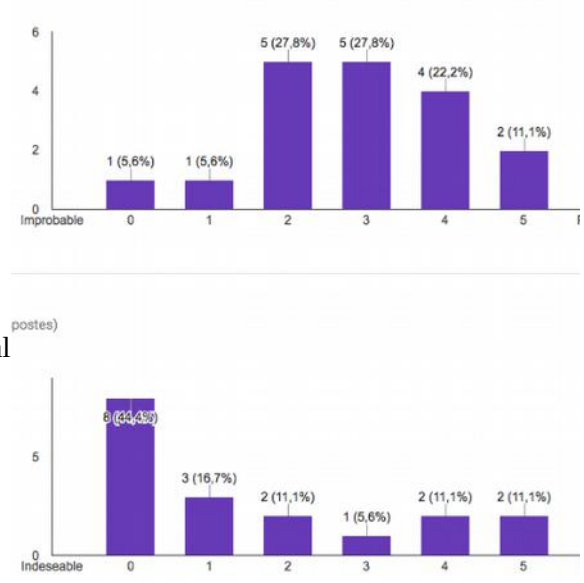


Image-18

Statement - 2 (Image-18)

"In the short term nothing will change in Social Sciences. Research in forecast of political trends in Social Media will be held, as today, mostly by researchers related to Computer Science or Mathematics".

Apparently, the experts are a bit skeptical about witnessing big changes in the profile of the researchers working in the field. The avg. point is just slightly positive 2.89 and roughly 40% do not expect any change. Which is perceived by almost anybody as negative (avg. point 1.56).

The next statements try to figure out why the researchers have this negative view.

Statements - 4 (Image-19)

"Universities will offer the students of Social Sciences a good knowledge of the tools and know-how needed to analyze the digital trail that people leave on the Internet".

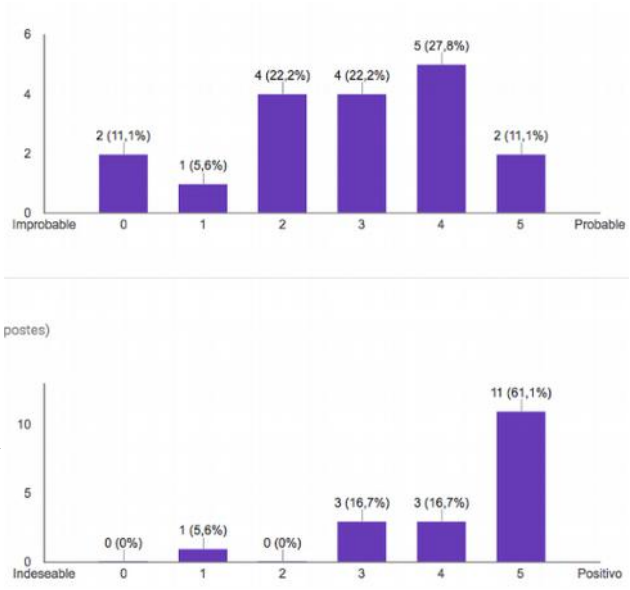


Image-19

The majority of experts foresee it as likely (more then 50% avg. point 2.83) that University will provide an appropriate formation and see it highly desirable (61%). In fact one expert says:

"*los investigadores en ciencias sociales ya utilizamos desde hace años herramientas informáticas, matemáticas y estadísticas*" Which is true, but maybe they do not feel comfortable enough with them in order to do research in the field. Academic literature shows that 90% of the papers in Big Data linked with Sociology or Political Science for political prediction were produced by Computer Science researchers.

So, experts agree that knowledge in those area offers a great enhancement to students and its possibilities will be more relevant soon, therefore Socials Science students will have to acquire them in order to do research. But maybe this is more of a wishful thinking that a real chance as we can see from next Statement.

Statement - 5 (image-20)

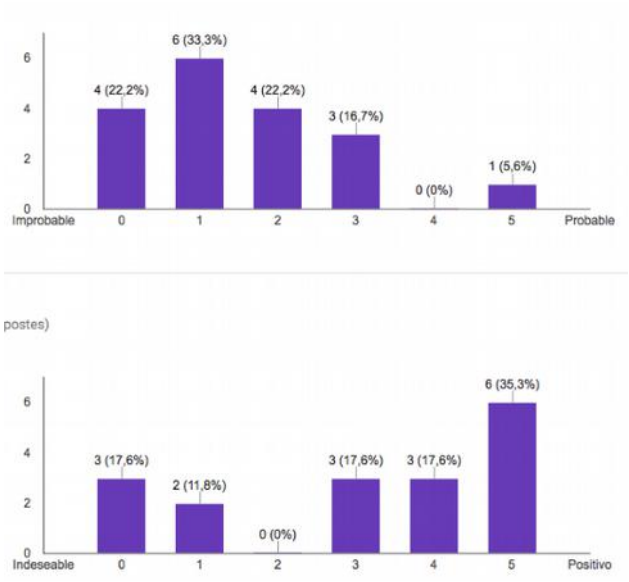


Image-20

"Social Science students' formation in Statistics and Computer Science is correct".

When the experts are as asked about the current situation they grade Spanish Social Science student's formation in Computer Science and Statistics as very bad, more than (75 %) believe that actually students do not have good enough skills in this areas (avg. 1.56). This may help to explain why when asked about the future of research they say that most probably the following years research will be done, as currently is done, by mathematicians and Computer Science scientists rather than the Social Science researchers.

So, although Sandra Gonzalez says:

"Social scientists need to embrace Big Data and gain literacy in the research that it makes possible, even if it goes beyond traditional disciplinary boundaries. The implications of that research for how we understand the social world are huge—and it is part of our remit to shape that understanding." (Gonzalez-Bailon, 2013)

According to the experts from the Delphi panel the current situation looks bad and doesn't look like it will change in the near future.

Statement - 6 (Image-21)

"It's not necessary for Social Sciences students' to learn Computer Science. Soon, tools will be developed to do the job of data mining, sorting algorithms, sentiment analysis, trend detection, etc. So students will be able to access the information datasets transparently".

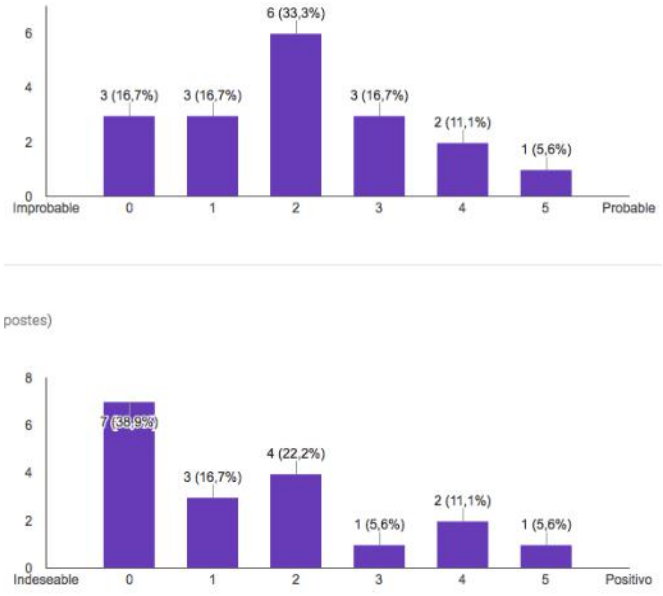


Image-21

What this statement express can be seen for some Social Science students as a very desirable point. If pretty soon applications will be created to do the job of digging the information datasets and sort the information without the need for researchers to know skills like data mining algorithms or classifiers then there is no need to learn those Computer Science related skills, they may have more time and resources to focus on their tools. The results show that, even if this applications are created, the majority of experts see it as unlikely that this tools will be useful (avg. point 2.06) and found it not desirable at all (38,9%) that this kind of tools may stop Social Science Students from learning the techniques needed for data mining, machine learning, sentiment classification, etc. As one experts points:

"Aun existiendo herramientas, es necesario cierto nivel de conocimientos sobre lo que se maneja"

Besides, when researchers use tools made by third parties sometimes it's difficult to know the bias or to replicate the experiments because those tools have disappeared. For instance Taratweet (http://taratweet.blogspot.com.es/) -a web application with which you can follow the Twitter social conversation based on a series of chosen hashtags- was used to do some research (*taratweet.blogspot.com)* but the application is not anymore online and Deltell (Deltell, Osteso, & Claes, 2013) used lots of third party applications in their paper which are not anymore available. If a researcher crates an algorithm or

types code, it can always be adapted or analyze, but third party applications are kind of black boxes with an unknown death-line.

Statement - 7 (Image-22)

"Researchers in Social Sciences will have to integrated into mixed teams of computer scientists and mathematicians as the level of knowledge to access and manage large databases will be too high".

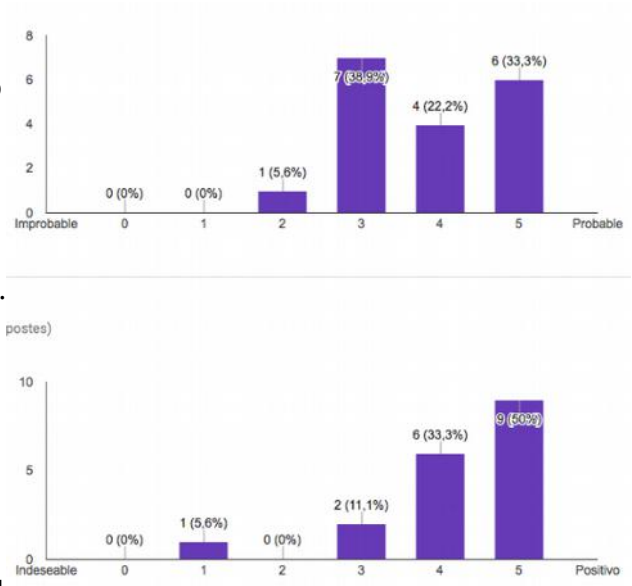


Image-22

Almost all the experts (more then 90 %) seem to believe that probable or most likely Social Science students will have to consider joining mixed research teams due to the increasing complexity and size of digital datasets. And again almost all of them consider this inter-field collaboration as very positive one. As Sandra Gonzalez says (Gonzalez-Bailon, 2013) says "social scientists can no longer do research on their own: the scale of the data that we can now analyze, and the methods required to analyze them, can only be developed by pooling expertise with colleagues from other disciplines".

Discussion

I have reviewed the academic literature in order to find out what it is already possible in data mining Twitter for political trends detections. I have highlighted the standard algorithms and the caveats and limits of data collection. I have also identified all the researchers that have published papers on this subject, their academic profiles and the key persons on this field of research. Whit this information I have created a list of experts for a Delphi panel and submitted to them a survey-interview in order to know the-state-of-the-art in this area and they perception of the near future.

Now, thanks to all the information acquired I will conclude this article enumerating briefly the main weaknesses of current research regarding political prediction with Twitter data and what is feasible to do nowadays. I will also provide a number of recommendations for future research (what is probable) and a desired scenario for Social Sciences; what is desirable that Social Sciences has to do to keep updated with data mining. I will revisit our research questions based on the results obtained, and see how they can be answered.

1 - Can methods like data-mining, sentiment analysis or similar be use in Spain more successfully then offline polls?

So far, the simplest answer is: it depends. It's still very complicated to deal with problems like normalization, language, lack of demographic information and different kind of bias. Lexicon based sentiment analysis and sorting techniques may improve the prediction result, but the improvement also vary in different states or regions. Probably it will be helpful for political trend detection whenever there just a couple of options for the voters to choose or when it affects a big part of the population which is not very geographically distinct. For instance, a referendum in Catalonia about the secession may work fine, but to detect the x per cent percent of a Catalan party in the Spanish parliament will most likely fail.

2 - Is there a need to modify Social Scientist's capabilities and skills (like Programing languages, statistics skills, networks knowledge) in order to adapt them to new technological environments?

Yes, the knowledge needed is crucial. Social Science researchers do not have to become Computer Scientists, but they do have to acquire basic programing capabilities and a good knowledge of Social Media, Internet and Big Data (Gonzalez-Bailon, 2013) (Schoen et al., 2013a) (Schoen et al., 2013b)

3 – Do we know what has to be improved in Twitter prediction in order to be at least as good as the off-line polls in Spain?

Yes. Mostly is necessary to tackle completely the drawbacks that Daniel Gayo (Gayo-Avello, 2012b) declares as the three major problems, that have to be addressed by future research:

1) the need to produce a true forecast, that is published before the elections.

2) the need to take into account the biases on Twitter, especially the unrepresentativeness of the sample.

3) the need to incorporate sentiment rather than just tweets volume (Sloan et al., 2015)

Furthermore there is an important need to create a standard methodological framework for data mining and a solid sociopolitical theoretical background in order to know the background logic between the metrics and the final prediction result. Researchers in most of the cases just use a collection of metrics to be trained on test data, find out which ones have the highest coefficients, and use them to compose the prediction model. Consequently, lacking a solid supporting theory, we cannot be sure that one model, which works well in one case, could be applied to other situations with the same accuracy (Yu & Kak, 2012). That's why some models show good performance in one election prediction, but completely fail in another one. To guarantee that one model has a good performance in all cases, we need to know the logic and theory behind the model. An excellent example of this problem is the recommendations case (Leskovec, Adamic, & Huberman, 2007) which proves that the recommendation on DVDs is more likely

to be accepted than that on books, but we still do not have a theory that may explain why these differences happens, and certainly the theory will not pop out just because Computer Science researchers and Mathematicians produce better algorithms. A humanistic or social approach is getting required in order to create a conceptual framework that allows research to improve on a solid basement.

4 – Are Spanish Social Scientist prepare to use those methods?

Although there are some papers about Twitter as a tool for political analysis produced by Spanish Social Scientist (for a full listing (Vilares, 2016)), when it comes to political forecast I found just one paper (Deltell et al., 2013), and it was an old and not really accurate one and another one that in a way it can be also considered (Soler et al., 2012) although it made correlations not predictions. All the other ones were either produced outside Spain or produced from Spanish Computer Science researchers.

The statistics, the network review and the results from the Delphi panel are quite clear, the Spanish Social Scientists are producing nothing in this field and the future is not very optimistic.

Conclusion

In this paper, I have presented a survey of political prediction using social media. I have offered an overview of the current research, the researchers profile and listed challenging problems and the areas for further research.

I have adopted and author centric-approach which complements the current concept-centric approaches and through this approach I have demonstrated that there is an enormous lack of Social Scientists doing research on how to use user generated digital data to predict political attitudes and trends.

There are some clear bias that have to be fixed, and there is an urgent need of a theoretical framework that could offer models which may suit society and people behavior. But experts believe that all this can be partly or fully overcome in the near future. Techniques for sentiment analysis or machine learning will improved, better datasets will be generated, more people will join the SNSs and more

accurate predictive knowledge will be acquired. Big corporations, public institutions, computer scientists and mathematicians are already working on this because they want to be part of this new deal and get some of its benefits. But the experts think that the knowledge and the education that Social Science students have nowadays it's not appropriate for them, so they can not to be part of this revolution neither going interdisciplinary teams.

Although political prediction using social media is only an emerging research topic and in most of the cases its results have so far relatively low accuracy, it has been proven that in some scenarios they can be extremely accurate.

I have shown that experts believe that the current theoretical framework and the techniques can be used in Spain with the Spanish Institutions. But there is almost no research going on in Spain. The only spanish scientist doing research in the field are working abroad and the few ones doing research in Spain are not from the Social Sciences field but from Computer Science. Indeed, there is a lot of research going on from the Computer Science area but almost none from the Social Sciences. And this is not positive at all. Like in the offline world, to forecast where people goes and to create tools to measure and control its flow is important, but without knowing why people moves and what makes them move, the research can not really advance, and certainly doesn't make much sense when it comes to better understand why things happen; which is science's core goal.

Social Science and Computer Science have to learn how to work together.

Appendixes

1 Appendix 0 - List of experts contacted for the Delphi panel.

1. Alonso, Miguel A. (Universidade da Coruña)
2. Balcells, Laia (Duke University)
3. Barberá, Pablo (NY University)
4. Becerra Fernandez, Adrian (empresa)
5. Blanco, Carlos (empresa)
6. Cotelo Moya, Juan Manuel (Universidad de Sevilla)
7. Cruz Mata, Fermin (Universidad de Sevilla)
8. Doval, Yerai (Universidade da coruña)
9. Enriquez, Fernando (Universidad de Sevilla)
10. Ferràs Hernàndez, Xavier (Esade)
11. Fondevila Gascón, Joan-Francesc (UDG/ UAO CEU)
12. Gayo-Avello, Daniel (Universidad de Oviedo)
13. Gómez Rodríguez, Carlos (Universidade da Coruña)
14. González-Bailón, Sandra (University Pennsylvania)
15. Hurtado, Lluis F. (Universitat Politècnica de València)
16. Martín-Dancausa, Carlos (Robert Gordon University)
17. Camacho, David (UAM)
18. Nafría, Ismael (Periodista)
19. Ortega, F.Javier (Universidad de Sevilla)
20. Pla, Ferran (UPV)
21. Planagumà i Valls, Marc (Bigdatabcn.com)
22. Rivero, Gonzalo (NYU)
23. Troyano Jiménez, Jose Antonio (Universidad de Sevilla)

24. Vilares Calvo, David (UDC)
25. Zubiaga, Arkaitz (University of Warwick)
26. Baeza-Yates, Ricardo (UPF)
27. Poblete, Barbara (Universidad de Chile)
28. Bravo-Marquez, Felipe (Waikato University)
29. Jorge Fabrega (CICS/UDD)
30. Mendoza, Marcelo (Universidad de Chile)
31. García-Gavilanes, Ruth (University of Oxford)
32. Gema Bello Orgaz (Universidad Autonoma de Madrid)
33. Mariluz Congosto (Universidad Carlos III)
34. Florencia Claes (Universidad Complutense Madrid)
35. Héctor D. Menéndez (UCL)
36. Pablo Aragón (Universitat Pompeu Fabra & Eurecat)
37. Kiko Llaneras (Universidad Politécnica de Valencia)
38. Luis Deltell-Escolar (Universidad Complutense Madrid)
39. Andreas Kaltenbrunner (Scientific Director Barcelona Media)

2 Appendix 1 - Letter send for the Delphi panel.

Muchas gracias por acceder a participar en este estudio.

Cada pregunta o afirmación tiene un doble eje; lo que considera usted probable que ocurra y si considera deseable o positivo que eso ocurra. Un valor de 0 significa que considera muy improbable que eso ocurra y un valor de 5 que lo considera muy probable. Lo mismo ocurre con el eje de positividad. Un 0 lo considera poco deseable y un valor de 5 muy positivo o deseable.

En cada afirmación o cuestión hay también un campo de texto opcional por si desea añadir cualquier tipo de información o matizar su respuesta. En el campo de texto puede utilizar el catalán, el castellano o el inglés. El trabajo será publicado en inglés por lo que todos los textos serán traducidos a ese

idioma, aunque en el apéndice final se incluirá el texto original en catalán o castellano de cualquier texto citado.

Las respuestas serán anónimas. Se le indica que incluya su nombre únicamente para identificar a efectos estadisticos el origen de las respuestas.

Las frases y el contexto hacen referencia a Twitter o la red de microblogging que le substituya.

a) Referente a la información en los mensajes (tweets)

1.- En breve será posible idear métodos para que los clasificadores utilizados para determinar el sentimiento o la ideología de los mensajes entiendan las sutilezas, giros e idiosincrasias del idioma y detecten spam, repeticiones, etc?

2.- En breve, gracias a diferentes técnicas de obtención de información o datasets externos, será posible eliminar los sesgos demográficos y obtener perfiles socio-demográficos precisos de los usuarios de Twitter

3.- En breve la muestra poblacional que las SNS faciliten no tendrá sesgos desconocidos

4.- En breve los gobiernos intentaran influir en la información del dataset que las Social Media faciliten a los investigadores

5.- En breve será posible acceder gratis a todos los datos de los usuarios de las Social Media

6.- En un breve aumentaran los costes de acceder a todos o la mayoría de los datos de los SNS

7.- En breve el interés comercial por las herramientas de gestión de minería de datos y “machine learning” provocará que en los papers cada vez más se utilicen «Black boxes» que oculten o no detallen partes de la investigación

8.- A medio plazo las uniones de Social Media, Telecos y Financieras creará datasets de usuarios muy precisos y estructurados no abiertos a los investigadores.

9.- Los grandes datasets públicos o privados pueden ser anonimizados y facilitados a los investigadores para mejorar la investigación u obtener conocimiento

b) Referente a las particularidades del sistema político Español

1.- Podemos trasladar fácilmente los marcos de trabajo utilizados en otros países a España

2.- Con la información disponible en Internet, en breve será posible hacer predicciones sobre los resultados de todos los partidos (independientemente de su tamaño o representatividad) con una MAE[6] igual o inferior a la de las encuestas electorales.

3.- Con la información disponible en Internet es posible hacer predicciones de los grandes partidos parlamentarios con una MAE igual o inferior a la de las encuestas electorales.

4.- Con la información disponible en Internet es posible hacer predicciones de grandes tendencias que impliquen a mucha gente y ofrezcan pocas opciones; (secesión - unión, monarquía - república, salida UE - permanencia, etc) con una MAE igual o inferior a la de las encuestas electorales.

5.- En España hay instituciones publicas o privadas con grandes datasets que pueden realizar análisis políticos y de tendencias políticos mucho más precisos que los que se publican

6.- En España hay instituciones públicas o privadas que comparten sus datasets con gobiernos o partidos políticos afines.

7.- Los gobiernos crearan nuevas instituciones para gestionar el conocimiento sobre comportamientos y tendencias sociales que es posible obtener gracias a la minería de datos

c) Referente a la formación de estudiantes e investigadores de ciencias sociales en España

1.- La necesidad de conocer herramientas de programación como R o Python se incrementará.

2.- La necesidad de conocer técnicas clasificadoras como SVM, Random Forest, Naïve Bayes, etc se incrementará en breve

3.- En breve nada cambiará en las ciencias sociales y la investigación en previsión de tendencias políticas se realizará, como ocurre hoy en día, sobretodo por parte de investigadores relacionados con la informática o las matemáticas

4.- A medio plazo, la potencia y las posibilidades de la minería de datos y las herramientas de «machine learning» será tan grande que se crearán nuevas especialidades en sociología, humanidades o ciencias políticas muy centradas en la gestión del rastro digital que las personas dejan en Internet.

5.- A los estudiantes de ciencias sociales se les facilitará en las universidades un buen conocimiento de las herramientas de gestión del rastro digital que las personas dejan en Internet.

6.- La formación en estadística e informática de los estudiantes de ciencias sociales es correcta

7.- No es necesario que los estudiantes de ciencias sociales aprendan informática pues en breve se crearán herramientas para que puedan acceder a la información de los datasets sin tener conocimientos de informática o minería de datos

6 Mean Absolute Error

8.- Los investigadores en ciencias sociales tendrán que integrarse en equipos mixtos de informáticos y matemáticos pues el nivel de conocimientos para acceder a grandes bancos de datos será demasiado alto.

3 Appendix 2 - Results of the Delphi Panel

	Nº Experto ->	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Media	Total m	Desvia
Tema 1.	Statement.1 - Var. A)	4	4	1	5	2	3	2	4	4	0	5	2	1	4	1	5	2	1	2,78	50	1,63
	Statement.1 - Var. B)	4	5	5	5	3	5	5	5	4	5	4	4	3	4	5	1	5	5	4,28	77	1,07
	Statement.2. -Var. A)	3	3	4	3	2	4	3	2	4	3	2	2	3	5	4	3	1	1	2,89	52	1,08
	Statement.2 - Var. B)	3	4	2	5	2	4	5	4	3	4	4	4	3	5	5	1	5	3	3,67	66	1,19
	Statement.3 - Var. A)	1	1	5	2	2	4	0	1	2	1	3	1	3	1	0	3	1	4	1,94	35	1,43
	Statement.3 - Var. B)	4	4	0	5	3	4	4	3	1	5	3	5	3	5	5	1	5	5	3,61	65	1,58
	Statement.4 - Var. A)	1	0	0	0	0	2	0	1	1	0	2	0	2	3	0	1	0	5	1,00	18	1,37
	Statement.4 - Var. B)	5	5	0	0	4	0	5	2	1	0	3	2	3	1	2	4	5	2	2,44	44	1,89
	Statement.5 - Var. A)	4	5	5	3	3	4	1	3	2	2	4	1	2	4	5	0	1	3	2,89	52	1,53
	Statement.5 -Var. B)	0	0	0	0	0	4	0	2	1	2	1	0	3	1	0	0	0	3	0,94	17	1,30
	Statement.6. - Var. A	4	1	1	1	1	3	0	3	1	0	3	1	4	3	2	4	3	3	2,11	38	1,37
	Statement.6 - Var. B)	0	0	0	4	0	3	0	2	1	0	3	1	2	0	0	0	2	0	1,00	18	1,33
	Statement.7. - Var. A	4	5	5	5	5	4	1	4	4	5	3	4	4	5	5	3	3	5	4,11	74	1,08
	Statement.7- Var. B)	1	0	0	4	1	1	0	2	0	1	3	2	1	0	0	0	2	0	1,00	18	1,19
	Statement.8. - Var. A	3	3	2	3	3	4	4	4	5	5	5	4	3	3	4	5	4	4	3,78	68	0,88
	Statement.8 - Var. B)	4	5	5	3	4	4	5	4	5	5	4	5	4	5	5	5	5	5	4,56	82	0,62
Tema 2.	Statement.1. - Var. A	3	5	3	3	5	4	3	2	4	3	2	4	4	1	3	2	1	3	3,06	55	1,16
	Statement.1 - Var. B)	4	5	5	5	5	5	5	5	4	4	1	4	4	4	5		5	5	4,41	75	1,00
	Statement.2. - Var. A	1	2	2	0	2	4	1	3	3	1	5	1	4	1	1	0	0	5	2,00	36	1,64
	Statement.2 - Var. B)	4	5	5	4	3	5	3	5	2	4	4	4	4	4	5	4	5	5	4,17	75	0,86
	Statement.3.- Var. A)	4	2	3	2	2	4	1	1	3	4	5	2	4	2	4	4	1	4	2,89	52	1,28
	Statement.3- Var. B)	5	5	5	4	3	5	3	4	2	5	4	4	4	4	5	5	5	5	4,28	77	0,89
	Statement.4. - Var. A	5	4	2	1	2	4	2	1	4	2	5	1	4	1	4	5	3	4	3,00	54	1,50
	Statement.4 - Var. B)	5	5	5	4	3	5	4	4	2	4	4	4	4	4	5	5	5	5	4,28	77	0,83
	Statement.5. - Var. A	4	1	0	0	2	3	2	2	1	4	5	2	3	4	5	5	2	4	2,72	49	1,64
	Statement.5 - Var. B)	2	1	0	2	0	4	4	4	0	2	3	3	2	1	3	0	2	2	1,94	35	1,39
	Statement.6. - Var. A	4	1	3	4	0	3	1	3	4	0	5	1	4	4	5	5	5	3	3,06	55	1,73
	Statement.6 - Var. B)	1	0	0	3	0	3	4	2	1	0	2	0	2	0	0	0	2	0	1,11	20	1,32
Tema 3.	Statement.1. - Var. A	4	4	4		4	4	3	4	4	5	3	5	4	4	5	5	5	3	4,12	70	0,70
	Statement.1 - Var. B)	4	5	3		4	5	5	5	4	5	3	5	4	3	5	5	5	3	4,29	73	0,85
	Statement.2. - Var. A	4	4	3	3	1	2	2	2	4	5	2	2	3	4	0	5	3	3	2,89	52	1,32
	Statement.2 - Var. B)	0	0	0	1	0	1	4	2	4	5	0	0	3	0	0	5	1	2	1,56	28	1,85
	Statement.3. - Var. A	4	5	4	5	5	4	4	3	3	5	3	4	2	5	5	5	4	4	4,11	74	0,90
	Statement.3 - Var. B)	3	5	5	5	5	5	5	3	4	5	4	4	2	5	5	4	5	5	4,39	79	0,92
	Statement.4. - Var. A	1	4	0	3	5	4	4	2	2	0	3	4	2	3	5	2	3	4	2,83	51	1,50
	Statement.4 - Var. B)	5	5	5	1	5	5	5	3	3	4	4	4	3	5	5	5	5	5	4,28	77	1,13
	Statement.5. - Var. A	1	3	2	0	1	3	0	1	2	0	0	5	3	1	1	2	1	2	1,56	28	1,34
	Statement.5 - Var. B)	4	3	3	0	1	5	5		4	0	5	5	3	5	0	1	5	4	3,12	53	1,96
	Statement.6. - Var. A	2	3	2	4	2	2	0	1	5	2	3	0	2	4	1	1	0	3	2,06	37	1,43
	Statement.6 - Var. B)	2	4	2	1	2	0	0	1	5	0	4	0	2	1	0	0	0	3	1,50	27	1,62
	Statement.7. - Var. A	3	3	5	3	4	3	3	4	5	3	2	3	4	5	5	5	5	4	3,83	69	0,99
	Statement.7 - Var. B)	3	4	4	5	5	4	5	4	5	3	1	4	4	5	5	5	5	5	4,22	76	1,06

References

Aharony, N., & Bronstein, J. (2014). A Delphi investigation into future trends in e-learning in Israel. *Interactive Learning Environments*, *22*(6), 789–803. http://doi.org/10.1080/10494820.2012.738232

Barbera, P. (2016). Less is more ? How demographic sample weights can improve public opinion estimates based on Twitter data . *Working Paper Para NYU*.

Barbera, P., & Rivero, G. (2011). Desigualdad en la discusión política en Twitter. In *Ponencia preparada para el I Congreso Internacional en Comunicación Política y Estrategias de Campaña* (pp. 1–22).

Barbera, P., & Rivero, G. (2014). Understanding the Political Representativeness of Twitter Users. *Social Science Computer Review*, *33*(6), 0894439314558836–. http://doi.org/10.1177/0894439314558836

Bermingham, A., & Smeaton, A. F. (2011). On Using Twitter to Monitor Political Sentiment and Predict Election Results. *Psychology*, 2–10.

Bollen, J., Pepe, A., & Mao, H. (2009). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *Www2010*, 17–21. Retrieved from http://arxiv.org/abs/0911.1583

Cameron, M. P., & Barrett, P. (2013). New Zealand Can Social Media Predict Election Results ? Evidence from New Zealand, *64*(0), 16. http://doi.org/10.1080/15377857.2014.959690

Castillo, C., Mendoza, M., & Poblete, B. (2013). Predicting information credibility in time-sensitive social media. *Internet Research*, *23*(5), 560–588. http://doi.org/10.1108/IntR-05-2012-0095

Ceron, A., Curini, L., & Iacus, S. M. (2014). Using social media to forecast electoral results. A meta-analysis. *UNIMI-Research Papers in …*, *25*(3). Retrieved from http://services.bepress.com/unimi/statistics/art62/

Choudhury, M. De, Lin, Y.-R., Sundaram, H., Candan, K. S., Xie, L., & Kelliher, A. (2010). How does the data sampling strategy impact the discovery of information diffusion in social media? *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, 34–41. http://doi.org/papers3://publication/uuid/8A1DA629-F17A-4F21-9FBD-4CAA6C0D32A7

Chung, J., & Mustafaraj, E. (2011). Can collective sentiment expressed on twitter predict political elections? *Aaai*, 1770–1771. http://doi.org/10.1007/s00247-002-0848-7

Compton, R., Jurgens, D., & Allen, D. (2015). Geotagging one hundred million Twitter accounts with total variation minimization. *Proceedings - 2014 IEEE International Conference on Big Data, IEEE Big Data 2014*, 393–401. http://doi.org/10.1109/BigData.2014.7004256

Conover, M., Ratkiewicz, J., & Francisco, M. (2011). Political polarization on twitter. *Icwsm*, *133*(26), 89–96. http://doi.org/10.1021/ja202932e

Deltell, L., Osteso, J., & Claes, F. (2013). Predicción de tendencia política por Twitter: Elecciones andaluzas 2012. *Ámbitos, Revista Internacional de Comunicación*, *22*, 13. Retrieved from http://www.researchgate.net/publication/236152207_Political_Trends_Prediction_on_Twitter_Andalusian_Election_2012__Prediccin_de_tendencia_poltica_por_Twitter_Elecciones_Andaluzas_2012/file/504635166e9b274682.pdf

Diaz, F., Gamon, M., Hofman, J. M., Kiciman, E., & Rothschild, D. (2016). Online and social media data as an imperfect continuous panel survey. *PLoS ONE*, *11*(1), 1–21. http://doi.org/10.1371/journal.pone.0145406

DiGrazia, J., McKelvey, K., Bollen, J., & Rojas, F. (2013). More tweets, more votes: Social media as a quantitative indicator of political behavior. *PLoS ONE*, *8*(11), 1–6. http://doi.org/10.1371/journal.pone.0079449

Duch, R. (2016). Tweet as a Tool for Election Forecast : UK 2015 General Election as an Example, 1–47.

Evangelos, K., Efthimios, T., & Konstantinos, T. (2014). Understanding the predictive power of social media. *Internet Research*, *23*(5). Retrieved from http://www.emeraldinsight.com/doi/full/10.1108/IntR-06-2012-0114

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, *1*(3), 215–239.

Gaurav, M., Srivastava, A., Kumar, A., & Miller, S. (2013). Leveraging candidate popularity on Twitter to predict election outcome. *Proceedings of the 7th Workshop on Social Network Mining and Analysis - SNAKDD '13*, 1–8. http://doi.org/10.1145/2501025.2501038

Gayo-Avello, D. (2012a). "I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper" -- A Balanced Survey on Election Prediction using Twitter Data, 13. Computers and Society; Computation and Language; Physics and Society. Retrieved from http://arxiv.org/abs/1204.6441

Gayo-Avello, D. (2012b). No, You Cannot Predict Elections with Twitter. *IEEE Internet Computing*, *16*(6), 91–94. http://doi.org/10.1109/MIC.2012.137

Gayo-Avello, D. (2013). *A Meta-Analysis of State-of-the-Art Electoral Prediction From Twitter Data. Social Science Computer Review* (Vol. 31). http://doi.org/10.1177/0894439313493979

Gayo-Avello, D., Metaxas, P. T., & Mustafaraj, E. (2011). Limits of Electoral Predictions Using Twitter - 3254. Retrieved March 2, 2016, from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2862/3254

Golder, S. A., & Macy, M. W. (2014). Digital Footprints: Opportunities and Challenges for Online Social Research. *Annu. Rev. Sociol*, *40*, 129–52. http://doi.org/10.1146/annurev-soc-071913-043145

Gonzalez-Bailon, S. (2013). Social science in the era of big data. *Policy and Internet*, *5*(2), 147–160. http://doi.org/10.1002/1944-2866.POI328

Gonzalez-Bailon, S., Borge-Holthoefer, J., Rivero, A., & Moreno, Y. (2011). The Dynamics of Protest Recruitment through an Online Network. *Scientific Reports*, *1*, 1–7. http://doi.org/10.1038/srep00197

Gonzalez-Bailon, S., Wang, N., Rivero, A., Borge-Holthoefer, J., & Moreno, Y. (2012). Assessing the Bias in Communication Networks Sampled from Twitter, *44*(0), 35. http://doi.org/10.2139/ssrn.2185134

Gordon, J. (2013). *Comparative Geospatial Analysis of Twitter Sentiment Data during the 2008 and 2012 US Presidential Elections*. University of Oregon.

Gruhl, D., Guha, R., Kumar, R., Novak, J., & Tomkins, A. (2005). The Predictive Power of Online Chatter. *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 78–87. http://doi.org/10.1145/1081870.1081883

He, W., Zha, S., & Li, L. (2013). International Journal of Information Management Social media competitive analysis and text mining : A case study in the pizza industry, *33*, 464–472. http://doi.org/10.1016/j.ijinfomgt.2013.01.001

Huertas, P. L., Moro, A. I., & López, F. J. M. (2005). Los Delphi como fundamento metodológico predictivo para la investigación en sistemas de información y tecnologías de la información (IS/IT). *Pixel-Bit: Revista de Medios Y Educación*, (26), 89–112. http://doi.org/10.1007/s13398-014-0173-7.2

Huertas, P. L., & Moro, A. I. (2006). Los Delphi Como Fundamento Investigación En Sistemas De Información Y Tecnologías De La Información (Is / It) Delphi Like As a Predictive Methodological Base for Research of Information Systems and Information Technology (Is / It). *Pixel-Bit: Revista de Medios Y Educación*, (26), 89–112.

Jernigan, C., & Mistree, B. F. T. (2009). Gaydar: Facebook friendschips expose sexual orientation. *First Monday*, *14*(10), 1–11. http://doi.org/10.5210/fm.v14i10.2611

Jungherr, A. (2014). Twitter in Politics: A Comprehensive Literature Review. *Available at SSRN*, 1–90. http://doi.org/10.2139/ssrn.2402443

Jungherr, A., Jurgens, P., & Schoen, H. (2011). Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. "Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment." *Social Science Computer Review*, *30*(2), 229–234. http://doi.org/10.1177/0894439311404119

Jurgens, P., & Jungherr, A. (2016). A Tutorial for Using Twitter Data in the Social Sciences: Data Collection, Preparation, and Analysis. *SSRN Electronic Journal*. http://doi.org/10.2139/ssrn.2710146

Jurgens, P., Jungherr, A., & Schoen, H. (2011). Small Worlds with a Difference : New Gatekeepers and the Filtering of Political Information on Twitter GATEKEEPERS DURING THE. *Websci' 11*, 1–5. http://doi.org/10.1145/2527031.2527034

Lampos, V., De Bie, T., & Cristianini, N. (2010). Flu detector - Tracking epidemics on Twitter. *Machine Learning and Knowledge Discovery in Databases*, 599–602. http://doi.org/10.1007/978-3-642-15939-8

Lampos, V., Preotiuc-Pietro, D., & Cohn, T. (2013). A user-centric model of voting intention from Social Media. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, 993–1003.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., … Al, Vanstyne, M. (2009). Life in the Networ: The Coming Age of Computational Social Science. *Science*, *323*(5915), 721–723. http://doi.org/10.1126/science.1167742.Life

Leskovec, J., Adamic, L. A., & Huberman, B. A. (2007). The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, *1*(1), 5.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167. http://doi.org/10.2200/S00416ED1V01Y201204HLT016

Mejova, Y., Weber, I., & Macy, M. W. (2015a). *Twitter: A Digital Socioscope*. Cambridge University Press. http://doi.org/10.1017/CBO9781316182635

Mejova, Y., Weber, I., & Macy, M. W. (2015b). *Twitter: A Digital Socioscope*. http://doi.org/10.1017/CBO9781316182635

Mishne, G., & Glance, N. (2005). Predicting Movie Sales from Blogger Sentiment. *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 155–158. http://doi.org/10.1016/j.cger.2010.02.002

Mislove, A., Lehmann, S., Ahn, Y., Onnela, J., & Rosenquist, J. N. (2011). Understanding the Demographics of Twitter Users. *Artificial Intelligence*, 554–557. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2816/3234

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. (2013). Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. *Proceedings of ICWSM*, 400–408. http://doi.org/10.1007/978-3-319-05579-4_10

Murthy, D. (2015). Twitter and elections: are tweets, predictive, reactive, or a form of buzz? *Information, Communication & Society*, *18*(7), 816–831. http://doi.org/10.1080/1369118X.2015.1006659

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. a. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series*, 122–129. http://doi.org/citeulike-article-id:7044833

O'Leary, D. E. (2015). Twitter Mining for Discovery, Prediction and Causality: Applications and Methodologies. *Intelligent Systems in Accounting, Finance and Management*, *22*(3), 227–247. http://doi.org/10.1002/isaf.1376

Ono, R., & Wedemeyer, D. J. (1994). Assessing the validity of the Delphi technique. *Futures*, *26*(3), 289–304. http://doi.org/10.1016/0016-3287(94)90016-7

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis Bo. *Foundations and Trends® in Information Retrieval*, *2*(2), 1–135. http://doi.org/10.1561/1500000001

Pennacchiotti, M., & Popescu, A. (2011). to Twitter User Classification. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media A*, 281–288.

Prasetyo, N. D. (2014). Tweet-Based Election Prediction, (December).

Schoen, H., Gayo-Avello, D., Takis Metaxas, P., Mustafaraj, E., Strohmaier, M., & Gloor, P. (2013a). The power of prediction with social media. *Internet Research*, *23*(5), 528–543. http://doi.org/10.1108/IntR-06-2013-0115

Schoen, H., Gayo-Avello, D., Takis Metaxas, P., Mustafaraj, E., Strohmaier, M., & Gloor, P. (2013b). The power of prediction with social media. *Internet Research*, *23*(5), 528–543. http://doi.org/10.1108/IntR-06-2013-0115

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., … Ungar, L. H. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*, *8*(9). http://doi.org/10.1371/journal.pone.0073791

Sloan, L., Morgan, J., Burnap, P., & Williams, M. (2015). Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. *PLoS ONE*, *10*(3), 1–20. http://doi.org/10.1371/journal.pone.0115545

Soler, J. M., Cuartero, F., & Roblizo, M. (2012). Twitter as a tool for predicting elections results. *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012*, 1194–1200. http://doi.org/10.1109/ASONAM.2012.206

Tan, S., Li, Y., Sun, H., Guan, Z., Yan, X., Bu, J., … He, X. (2014). Interpreting the public sentiment variations on Twitter. *IEEE Transactions on Knowledge and Data Engineering*, *26*(5), 1158–1170. http://doi.org/10.1109/TKDE.2013.116

Tsakalidis, A., Papadopoulos, S., Cristea, A., & Kompatsiaris, Y. (2015). Predicting the EU 2014 Election Results in Multiple Countries Using Twitter. *Intelligent Systems, IEEE*, *30*(March), 10–17.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 178–185. http://doi.org/10.1074/jbc.M501708200

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2011). Election Forecasts With Twitter: How 140 Characters Reflect the Political Landscape. *Social Science Computer Review*, *29*(4), 402–418. http://doi.org/10.1177/0894439310386557

Vilares, D. (2016). A review on political analysis and social media ∗. *Procesamiento de Lenguaje Natural*, (56), 13–23.

Watts, D. J. (2011). *Everything is obvious once you know the anwser*. (Crown Publishing Group, Ed.). Crown Publishing Group.

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a. *MIS Quarterly*, *26*(2), 13–23. Retrieved from http://www.jstor.org/stable/4132319

Wlodarczak, P., Soar, J., & Ally, M. (2015). What the Future Holds for Social Media Data Analysis, *9*(1), 16–19.

Wong, F. M. F., Sen, S., & Chiang, M. (2012). Why Watching Movie Tweets Won't Tell the Whole Story? *Proceedings of the WOSN Conference*, (c), 1–6. http://doi.org/10.1145/2342549.2342564

Xie, Y., Yu, F., Ke, Q., Abadi, M., Gillum, E., Vitaldevaria, K., … Mao, Z. M. (2012). Innocent by Association : Early Recognition of Legitimate Users. *ACM Conference on Computer and Communications Security (CCS '12)*, 353–364. http://doi.org/10.1145/2382196.2382235

Yu, S., & Kak, S. (2012). A survey of prediction using social media. *arXiv Preprint arXiv:1203.1647*, 1–20. Retrieved from http://arxiv.org/abs/1203.1647