Opinion Mining in Social Networks versus Electoral Polls

Javi Fernández, Fernando Llopis, Yoan Gutiérrez, Patricio Martínez-Barco, Álvaro Díez
Department of Software and Computing Systems, University of Alicante
{javifm,llopis,ygutierrez,patricio,adiez}@dlsi.ua.es

Abstract

The recent failures of traditional poll models, like the predictions in United Kingdom with the Brexit, or in United States presidential election with the victory of Donald Trump, have been noteworthy. With the decline of traditional poll models and the growth of the social networks, automatic tools are gaining popularity to make predictions in this context. In this paper we present our approximation and compare it with a real case: the 2017 French presidential election.

1 Introduction

Numbers leave no doubt: *social networks* are becoming more popular each day. According the *Internet Live Stats* website[1], around 46% of the world population has an Internet connection today (July 2017), and around 55% of them are active users in a social network like *Facebook*[2] or *Twitter*[3]. This means around 2 billion people using social networks in the world. Beyond the sociological information they can bring, social networks have become a place where users not only can learn about what is happening around them, but also give opinions with respect to their environment. From a political point of view, they provide a lot of information on the possible degradations of the standard of living of a particular geographical area (Intagorn and Lerman, 2013). Others prefer to conduct studies of more reliable sources such as the news of the digital press (Leetaru et al., 2013).

Despite the slight decline in usage compared to previous years, *Twitter* still has an important use among Internet users, with more than 500 million tweets per day. But what allows Twitter to be one of the most popular data sources for social research is its open API[4] (Leetaru et al., 2013). It offers an *"unprecedented opportunity to study human communication and social networks"* (Miller, 2011). The access to the information turned over Twitter by millions of users can offer us a snapshot of the general state of each location. This information can help with the early detection of conflicts. In addition, if these studies are carried out in real-time, in the event of a catastrophe we can obtain critical information about in which areas it is a priority to act.

One of the possible uses of this information provided in real-time is the analysis of opinions of the citizens about the candidates to some electoral campaign, especially because the classic models of electoral polls are in crisis. The recent failures of these traditional models, like the predictions in United Kingdom with the Brexit, or in United States presidential election with the victory of Donald Trump, are very well-known. In a world in which information grows faster and faster, it is striking that models based on making hundreds of phone calls are still employed to predict results, like the surveys in the *Financial Times*[5] (see Figure 1). Traditional models obtain information that will be obsolete before it can be published or even processed. This obsolescence is not only due to the cost of information processing, but also due to people who are becoming more and more tired of talking to pollsters, and in many cases their intention to vote reflects a timely opinion, that may change during the electoral process.

Social networks allow to measure opinions and emotions of the citizens throughout the whole

[1] www.internetlivestats.com
[2] www.facebook.com
[3] twitter.com
[4] Application Programming Interface
[5] ig.ft.com/sites/france-election/polls

Proceedings of Recent Advances in Natural Language Processing, pages 231–237,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_032

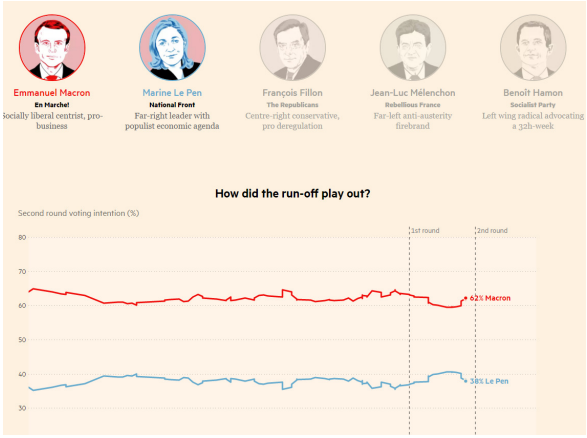

Figure 1: French election polls, Financial Times

electoral process. Users write their opinions about one candidate or another continuously. It is common to use the number of tweets or the *trending topics* as metrics to evaluate the reputation. But speaking a lot about someone does not imply the comments are necessarily positive. This is where the use of *Natural Language Processing* (NLP) and *Sentiment Analysis* (SA) techniques can be very useful, allowing to detect, with a reasonable degree of reliability, if the opinions about something are positive or negative.

In most of the cases, the users can provide information about their usual location or the place they write their messages. This allows to create studies with far more detail and information than traditional ones. However, location is not usually shared because of privacy or battery consumption. Some studies (Weidemann, 2014; Cebeillac and Rault, 2016) showed that the number of geolocated tweets is usually between 10% and 20%, but in our experience this number can be much lower, less than 5% in some contexts like politics. So, except some studies that use geolocation data to filter out unwanted messages (Poblete et al., 2011), most of them indicate that the users location is a field that determines with higher precision the geographical position of an user, and therefore their environment.

In this paper we present the *Social Analytics* system, developed by the *Natural Language Processing Group* (GPLSI) at the *University of Alicante*, a system that creates real-time reports summarising how different entities are being valued, based on the number of *positive*, *negative* or *neutral* opinions that users write about it, and the *audience* of those opinions (the number of followers of their authors). In addition, our system provides charts, aggregations, statistics, and advanced filters to show the information desired by the users. In previous works we studied the reputation in different contexts, but in this paper we present a study of the quality of the system to predict electoral results, and we compare it with a traditional model based on surveys. The context of experimentation chosen is the **2017 French presidential election**.

The article is structured as follows. In Section 2 we briefly describe the related work. Section 3 describes the architecture of our system. Section 4 will detail the formulas used to determine how the reputation of an entity is measured. Section 5 describes the evaluation, and Section 6 presents a series of conclusions and works that are currently being done to improve the system.

2 Related Work

Some other systems to visualize data from social networks exist, the majority of them focused only in statistics, but some others also adding semantic features as sentiment analysis (Marcus et al., 2011; Hao et al., 2011; Wang et al., 2012). There are also some public tools with this purpose, such as *SocialMention*[6]. In addition, there have been several studies in calculate the reputation of an entity, this is, how an entity is being valued in the Internet (Villena-Román et al., 2012; Amigó et al., 2014). Our proposal contains different statistical visualisations, reputation calculation, and advanced filtering by different dimensions, everything in real-time. The main goal of our system is to visualise the current state of an entity, what is being said about it, how it is being valued and, in some cases, make predictions about the future reputation of that entity.

It is essential to detect the polarity of the messages: if the data is expressing an opinion and, if it does, indicate if it is positive or negative. The field of *sentiment analysis* and *polarity classification* has been widely studied in the last years (Pang et al., 2008; Liu and Zhang, 2012; Mohammad, 2015; Ravi and Ravi, 2015). Two main approaches can be followed: *machine learning* and *lexicon-based*. *Machine learning* approaches perform very well in the domain they are trained on, but their performance drops when the same classifier is used in a different domain. In addition, if the number of features is big, the efficiency drops dra-

[6] www.socialmention.com

matically. *Lexicon-based approaches* make use of dictionaries of opinionated words and phrases to discern the polarity of a text. These approaches are usually faster than machine learning ones, as they employ predefined mathematical functions, but less effective in specific domains because they are (in general) more generic. The approximation we chose for our system is a hybrid approach, simple and fast enough to be used in real-time applications, but with a decent classification quality. This approach is described in Section 3.

Furthermore, in the case of Twitter, the usual location of the user is a free text field, so in some cases users can fill it with misspellings or an incorrect location. Obtain a real place from this field is a complex problem (Hecht et al., 2011; Peregrino et al., 2013). Our approximation to this problem is described in detail also in Section 3.

3 Architecture

In summary, our approximation downloads messages and comments from the social networks, extracts the useful information they contain (text, author, polarity, locations, etc.), and stores it in an efficient way to generate reports in real-time. We divided the system in three main modules: *listening*, *processing* and *presentation*, and uses three different databases (see Figure 2).

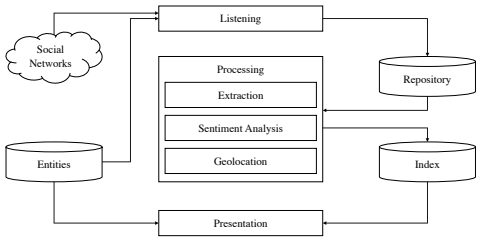

Figure 2: System architecture

3.1 Databases

First, we will describe the databases employed in our system. This will allow the reader to understand better how the concepts are represented and the information we want to fetch.

- **Entities**. In this database we store the entities of interest for the users. Users must manually assign terms (words, phrases, hashtags, or usernames) for each entity, that must appear in a text to consider it is mentioning that entity. For example, in the context of the 2017 French presidential election, the entity *Emmanuel Macron* would contain terms like "emmanuel macron", "macron" or @emmanuelmacron.

- **Repository**. This is a temporary database in which we store the messages and comments found in social networks. Here, all the data is stored without any processing, and is deleted once processed.

- **Index**. This is also a messages and comments database. But here, the data is stored in a way that permits to perform analytics and statistics as efficiently as possible. We want to highlight that the data is only indexed and not stored, so it is not accessible anymore. We can only perform statistics, aggregations and filters.

3.2 Listening module

This module is responsible for downloading messages from the social networks. It does periodic search requests of the terms in the entities database, using the social network API (when available) to fetch the messages containing those terms. The frequency of the periodic searches depends on the API limits and the number of terms we have in the database. Some APIs allow to obtain these messages in streaming, that is, offering the messages as soon as they are published. In both cases, the fetched messages are stored in the *Repository* database.

3.3 Processing module

This module performs the data extraction, location detection and sentiment analysis of the fetched messages through the listening module. When available, we extract the text of the message, its publication date, its author, the users that are mentioned, the location where the message was written, and the usual location of the author.

In the case of Twitter, this last field is free text and users can write whatever they want. To obtain the most probable location we employ a very simple approximation. We indexed a places database (*Geonames*[7]), performing *exact* searches on it, and choosing the first result. We tried with more relaxed queries, but we preferred to obtain better precision and avoid false positives.

[7] http://www.geonames.org

From the text we also extract its polarity, that is, if the author is giving a positive, negative or neutral opinion about the entity. To detect this polarity we employ a hybrid supervised approach (Fernández et al., 2013, 2014a,b, 2015; Gutierrez et al., 2015). In summary, this approach builds a sentiment lexicon from a polarity dataset using statistical measures. It uses *skipgrams* as information units, to enrich the sentiment lexicon with combinations of words that do not appear explicitly in the text. The lexicon created is employed in conjunction with machine learning techniques to create the final classifier. This approach has been chosen mainly because of its balance between speed and accuracy.

3.4 Presentation module

This module refers to the user interface. Figure 3 shows the main view of the dashboard in our system. At a glance we can access all the data from different points of view:

- The number of mentions for each entity.
- The audience[10] (number of people) the messages of each entity can potentially reach.
- The entity reputation, a numeric value representing how users are valuing the entity (see Section 4).
- The more repeated words and hashtags.
- The most active users with the biggest number of publications, and the most mentioned users (usually the most replied users).
- The places the messages were written and the geographical origin of the authors of the messages.
- The polarity of the messages.

One of the important features of our system is that all the information shown can be filtered, in real-time, by date, polarity, author, location, etc.

4 Reputation

Using the number of positive, neutral and negative opinions detected, and the audience of those opinions, we can calculate a value of **reputation** for each entity in a specific time period. This metric is a numeric value in the interval $[-1, +1]$, where -1 is the worst value of reputation and $+1$ is the best one. It is calculated using the formula in Equation 4, where e is the entity we are assessing; t is a time interval; $P_{e,t,+}$, $P_{e,t,0}$ and $P_{e,t,-}$ represent the set of publications containing a mention to the entity e in the time interval t with a positive, neutral and negative polarity respectively; a_p is the audience of the publication p; and d_t is the duration of the time interval t in milliseconds. Other statistics like the number of likes, retweets or clicks are not considered in this formula, mainly because they are not available at publication time (they are usually zero).

The inclusion of the time interval in the equation (d_t) is a way to give a higher reputation to those publications with a higher audience. For example, if an entity has negative mentions and reached 100 people in a minute, its reputation in that interval would be $-3 \cdot 100/(3 \cdot 100 + 60000) = -0.005$. However, if those publications reached 10,000 people, the reputation would be $-3 \cdot 10000/(3 \cdot 10000 + 60000) = -0.333$.

We want to highlight that in our system we consider neutral mentions ($p \in P_{e,t,0}$) as something positive, because being mentioned in social networks improves the reputation. In this way, if all the mentions are neutral, the reputation will be bigger than zero, while if there are no mentions, the reputation would be exactly zero.

$$pos = \sum_{p \in P_{e,t,+}} a_p \quad (1)$$

$$neg = \sum_{p \in P_{e,t,-}} a_p \quad (2)$$

$$neu = \sum_{p \in P_{e,t,0}} a_p \quad (3)$$

$$r_{e,t} = \frac{2 \cdot pos + neu - 2 \cdot neg}{2 \cdot (pos + neu + neg) + d_t} \quad (4)$$

In the context of the politics, this value of reputation between -1 and $+1$ may not be the most intuitive, because it is more common to use the *vote intention* measure. This is a percentage in the interval $[0\%, 100\%]$, where all the candidates sum 100% in every period of time. We made some additions in order to adapt our metrics to obtain a similar value. The formula for this new **collective reputation** is shown in Equation 5, where E is the set of all the entities to evaluate; and i is an entity inside the set E.

[10] This value is calculated by adding the number of followers of the authors of all the messages.

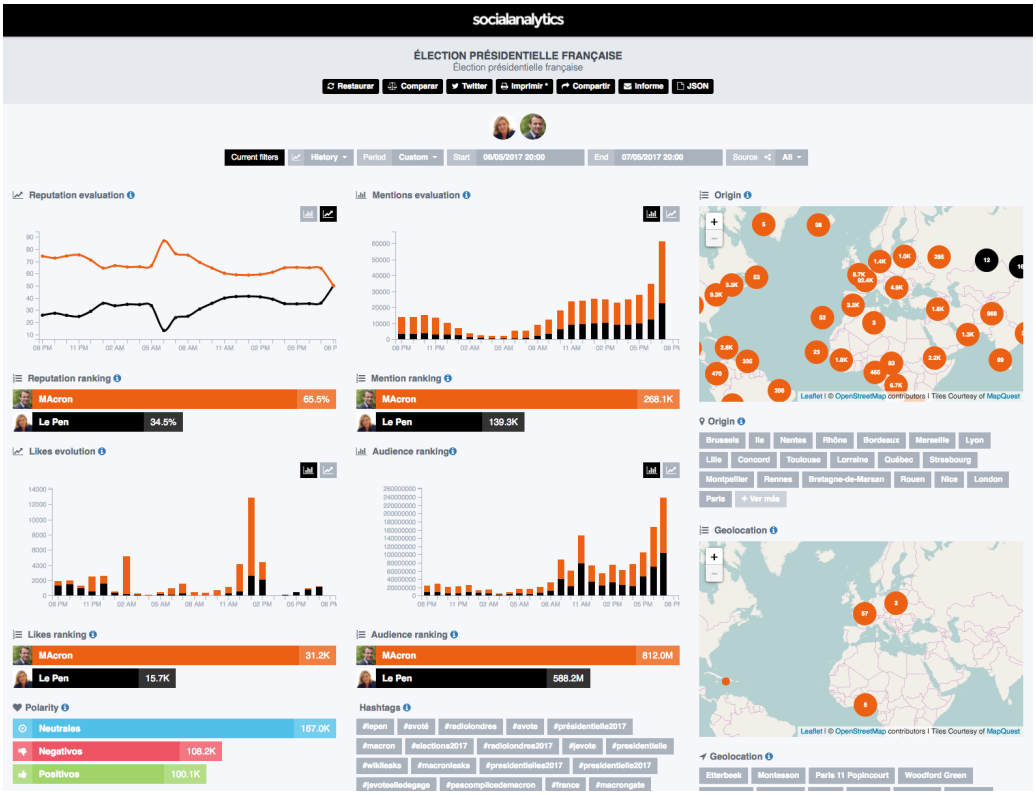

Figure 3: Main system dashboard, comparing the two main candidates in the French presidential election between May 4, 2017 and May 5, 2017 (Macron: orange, Le Pen: black)[9]

$$cr_{e,t} = \frac{r_{e,t} + 1}{\sum_{i \in E} (r_{i,t} + 1)} \cdot 100 \qquad (5)$$

This collective reputation is the metric we use in Figure 3 and the one we will use in the following Evaluation section.

5 Evaluation

The evaluation was performed in the context of the 2017 French presidential elections, specifically for the second round and the two most popular candidates: Emmanuel Macron and Marine Le Pen. In our opinion, this context was a suitable field of experimentation for the predictive use of our system, because:

- The circumscription was unique.
- There were only two candidates to evaluate.
- There was a lot of activity for both candidates, so there was a big number of messages to evaluate.

A week before the elections, we created an entity for each candidate. The terms chosen for Macron were "macron", "emmanuel macron", #macron, and @emmanuelmacron. For Le Pen, the terms were "le pen", "marine lepen", #lepen, and @lepen. Some days after, two predictions of election results were published on Twitter[11] (see Figure 5) (May 5, 2017 and May 6, 2017) the election day (May 7, 2017) using the proposed collective reputation. Some hours after the publication of our first prediction, a French television channel (BFMTV) published a poll that showed values[12] that were very similar to the ones given by our system (see Figure 4). However, our predictions were obtained in real-time, were easier to obtain, and were more economic.

During that weekend, the value of reputation for Macron was increased reaching the values of our second prediction[14] (see Figure 5). The final result was 66% for Macron and 33% for Le Pen. In our first consideration we concluded that the positive tendency of Twitter was not reflected in the final results. It is possible that if the election would be celebrated later, the result could not be

[11] twitter.com/fernandollopis/status/860463546597093376
[12] https://goo.gl/bzlwX1
[14] twitter.com/fernandollopis/status/860985783095885825

Figure 4: Vote intention published by BMFTV

the same. But what seems evident is that tendency of increase in Twitter.

There are other sociological aspects outside our study, but interesting to remark:

- We do not know the representation of the active French population with respect the total voting population.
- The influence of abstention in the final result.
- The information propagation speed and tendencies on the Internet respect the real vote propagation.

6 Conclusions

In this paper we presented the *Social Analytics* system. This system creates real-time reports summarising how different entities are being valued, providing a value of reputation for each entity. This reputation can be used to make comparisons with other entities and, in some cases, it can be useful to make predictions based on the current reputation of those entities.

The context of experimentation chosen was the second round of the 2017 French presidential election. In this simple case, with two candidates of electoral unique circumscription, the results prediction and tendencies in social networks have been very close to the final results, or at least as correct as the traditional polls. This opens a great amount of possibilities with respect the use of social networks, much more economic and updated, to measure the status of the candidates in electoral campaigns.

We open two research lines. The first is to continue experimenting with the reputation formula to obtain better and more accurate results. We will continue studying the concepts of number of mentions and audience, and how they affect each other. The second goal is to improve the location detection of the authors, in order to be more accurate in geographic level results.

Acknowledgements

This research work has been partially funded by the University of Alicante, Generalitat Valenciana, Spanish Government, Ministerio de Educación, Cultura y Deporte and Ayudas Fundación BBVA a equipos de investigación científica 2016 through the projects TIN2015-65100-R, TIN2015-65136-C2-2-R, PROMETEOII/2014/001, "Plataforma inteligente para recuperación, análisis y representación de la información generada por usuarios en Internet" (GRE16-01) and "Análisis de Sentimientos Aplicado a la Prevención del Suicidio en las Redes Sociales" (ASAP).

References

Enrique Amigó, Jorge Carrillo-de Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Edgar Meij, Maarten de Rijke, and Damiano Spina. 2014. Overview of replab 2014: author profiling and reputation dimensions for online reputation management. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, pages 307–322.

Alexandre Cebeillac and Yves-Marie Rault. 2016. Contribution of geotagged twitter data in the study of a social group's activity space. the case of the upper middle class in delhi, india. *Netcom. Réseaux, communication et territoires* (30-3/4):231–248.

Javi Fernández, José M Gómez, and Patricio Martínez-Barco. 2014a. A supervised approach for sentiment analysis using skipgrams. In *11th International Workshop on Natural Language Processing and Cognitive Science (NAACL)*.

Javi Fernández, Yoan Gutiérrez, José M Gómez, and Patricio Martınez-Barco. 2014b. Gplsi: Supervised sentiment analysis in twitter using skipgrams. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. pages 294–299.

Javi Fernández, Yoan Gutiérrez, José Manuel Gómez, Patricio Martínez-Barco, Andrés Montoyo, and Rafael Munoz. 2013. Sentiment analysis of spanish tweets using a ranking algorithm and skipgrams. In *XXIX Congreso de la Sociedad Espanola de Procesamiento de Lenguaje Natural (SEPLN 2013)*. pages 133–142.

Javi Fernández, Yoan Gutiérrez, David Tomás, José M Gómez, and Patricio Martínez-Barco. 2015. Evaluating a sentiment analysis approach from a business point of view .

Figure 5: French presidential election prediction[13]

Yoan Gutierrez, David Tomas, and Javi Fernandez. 2015. Benefits of using ranking skip-gram techniques for opinion mining approaches. In *eChallenges e-2015 Conference, 2015*. IEEE, pages 1–10.

Ming Hao, Christian Rohrdantz, Halldór Janetzko, Umeshwar Dayal, Daniel A Keim, Lars-Erik Haug, and Mei-Chun Hsu. 2011. Visual sentiment analysis on twitter data streams. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*. IEEE, pages 277–278.

Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H Chi. 2011. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, pages 237–246.

Suradej Intagorn and Kristina Lerman. 2013. Mining geospatial knowledge on the social web. *Using Social and Information Technologies for Disaster and Crisis Management* pages 98–112.

Kalev Leetaru, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Eric Shook. 2013. Mapping the global twitter heartbeat: The geography of twitter. *First Monday* 18(5).

Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining text data*, Springer, pages 415–463.

Adam Marcus, Michael S Bernstein, Osama Badar, David R Karger, Samuel Madden, and Robert C Miller. 2011. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, pages 227–236.

Greg Miller. 2011. Social scientists wade into the tweet stream. *Science* 333(6051):1814–1815.

Saif M Mohammad. 2015. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. *Emotion measurement* pages 201–238.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2(1–2):1–135.

Fernando S Peregrino, David Tomás, and Fernando Llopis. 2013. Every move you make i'll be watching you: geographical focus detection on twitter. In *Proceedings of the 7th Workshop on Geographic Information Retrieval*. ACM, pages 1–8.

Barbara Poblete, Ruth Garcia, Marcelo Mendoza, and Alejandro Jaimes. 2011. Do all birds tweet the same?: characterizing twitter around the world. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, pages 1025–1030.

Kumar Ravi and Vadlamani Ravi. 2015. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems* 89:14–46.

Julio Villena-Román, Sara Lana-Serrano, Cristina Moreno, Janine García-Morera, and José Carlos González Cristóbal. 2012. Daedalus at replab 2012: Polarity classification and filtering on twitter data. In *CLEF (Online Working Notes/Labs/Workshop)*. volume 60.

Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. 2012. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*. Association for Computational Linguistics, pages 115–120.

Chris Donald Weidemann. 2014. *Geosocialfootprint (2013): Social media location privacy web map*. University of Southern California.