2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8311

Relevance as an enhancer of votes on Twitter[1]

Arroba Rimassa, Jorge [a]; Llopis, Fernando [b] and Muñoz Guillena, Rafael [b]
[a] Universidad Central del Ecuador, Ecuador, [b] Department of Languages and Computer Systems, University of Alicante, Spain.

Abstract

The concept of the influence of Katz and Lazarfeld given in the last century has evolved thanks to the appearance of Social Networks and especially Twitter. Because this microblogging has allowed candidates for any election process to be closer to their electors and also allows an analysis of the contents of the messages to determine their polarity.

The relevance of the messages that measure the level of influence that can be had in the voters, incorporated into the traditional analysis of the Social Networks allow to have a greater degree of precision in the electoral predictions that are made using natural language processing, NLP.

We have introduced in the methodology that we propose a mechanism to enhance the votes of those messages that have a greater relevance and turn them into votes in order to improve the predictability of the electoral results.

The proposed methodology was applied in the election for President of the Republic of Ecuador that was held on February 19, 2017, obtaining a Mean Average Error, MAE = 1.4 that demonstrates the relevance of incorporating the variable Relevance as an enhancer of votes.

Keywords: *Relevance, Twitter, Election process.*

[1] This research work has been partially funded by the University of Alicante, Generalitat Valenciana , Spanish Government, Ministerio de Educación, Cultura y Deporte and ASAP - Ayudas Fundación BBVA a equipos de investigación científica 2016(FUNDACIONBBVA2-16PREMIO) through the projects, TIN2015- 65100-R, TIN2015-65136-C2-2-R, PROMETEOII/2014/001, GRE16- 01: "Plataforma inteligente para recuperación, análisis y representación de la información generada por usuarios en Internet" and "Análisis de Sentimientos Aplicado a la Prevención del Suicidio en las Redes Sociales" (PR16_SOC_0013).

Editorial Universitat Politècnica de València

1. Introduction

When Katz and Lazarsfeld (1955) formulate the notion of influence in *Personal Influence* in which they state that what matters is knowing three elements: *The audience*: knowing how many and how are the people who attend a message; *Content Analysis*: comprising the concept of the messages issued and the *Effect Analysis*: the impact of the mass media used; to understand the context and the conditions in which the "campaigns" were carried out in the media to modify the opinions and behaviors, they would never suppose that the voters, nowadays, would be totally communicated with their candidates, their messages are personal.

How this was achieved, by the appearance of Social Networks; politicians begin to interact with voters directly, receiving positions of unconditional acceptance and also rejection and repudiation of others. Is that Social Networks allow everyone to comment and especially as I would say (Eco, 2015). For the ease that the Twitter gives we will use this microblogging to make electoral predictions.

2. Related jobs

Since the appearance of Twitter in 2006, some authors have made predictions or post-processing of the results of electoral processes in the world using the information they can download from it. We have selected a sample of various electoral processes in the world in which using mentions and sentiments analysis, (Ceron, 2015) and (Singh, 2018) were used and in 45% of these works they used only the mentions method, which consists of counting the total number of downloads for one or the other candidate and 55%, used some classification method to determine the polarity of the messages, the most used being Naive Bayes. In 2012 there was the greatest increase in the predictive use of Twitter, reaching 36%, as shown in Figure 1.

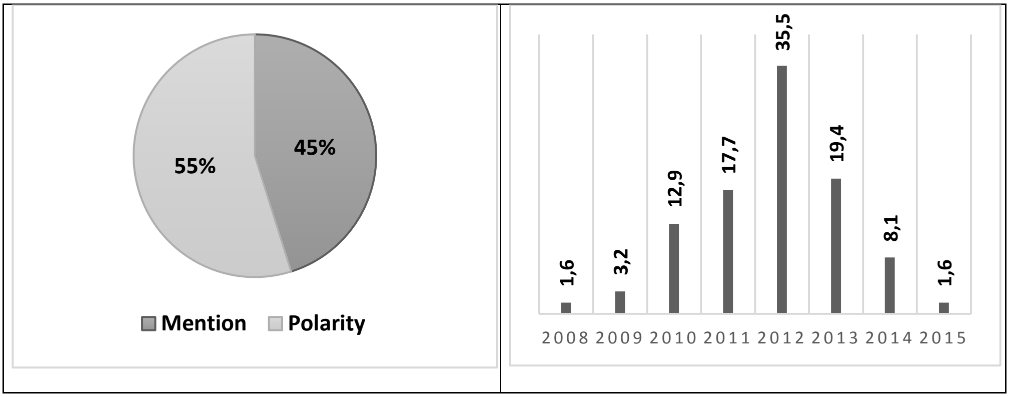

Figure 1. Method used to analyze and percentage of use for year. Source: Authors (2018).

3. Methodology

In order to predict the results of an electoral process using the analysis in the Social Networks, the opinion of the users in the network must be evaluated on both sides and the problem in question is focused on analyzing the position of each elector that the has manifested through a tweet and a way to solve this issue is to use natural language processing and text analysis to extract information. It must be analyzed how people interact (Blasquez & Domenech 2017).

The objective of the present investigation is to develop a methodology in which the concept and use of the Relevance as an enhancer of the votes of the messages that have a high influence are intended to be implemented.

As a practical case of application of this methodology, we used the Election for President of the Republic of Ecuador that was held on February 19, 2017.

The methodology developed is schematized in Figure 2. We will develop the steps of the methodology.

3.1. Defining accounts to follow

The accounts from which the documents were to be downloaded were first defined, the candidates Moreno, Lasso, Viteri and Moncayo were competing with some option for the first places for the presidency and a small percentage of the electorate would be for any of these others four candidates: Bucaram, Espinel, Zuquilanda and Pesántez; to these four candidates, for practical purposes they will be referred to as "others". In this sense, the official Twitter accounts of the candidates and their party were defined to extract the tweets from the users who follow them.

3.2. Twitter APP

Using the Twitter APP tool, the download was started from December 2016 until February 14, 2017; following the observations of (Tumasjan, 2010) on the periodicity of data collection. The elections would be on February 19, 2017 and a total of 823,135 tweets were downloaded corresponding to the users who follow these official accounts.

3.3. Collection of documents for each candidate

By gathering all these documents, the collection of documents was obtained for each of the candidates to be evaluated.

3.4. Pre-processing of Messages

This phase has the basic objective of reducing the dimensionality that usually presents problems later when the supervised learning methods of text classification are applied. The problems are not only due to computational reasons but also to the overfitting that can be presented in the dataset.

The mechanisms used for all the methods were: elimination of stopwords that contribute nothing to the text, the stopwords that were used are those that come by default in R. The cleaning process of blank spaces, markups, was also performed the emoticons, the reference links and the tags present in each text. Those terms whose distribution of frequency of appearance in the documents was much reduced were also eliminated.

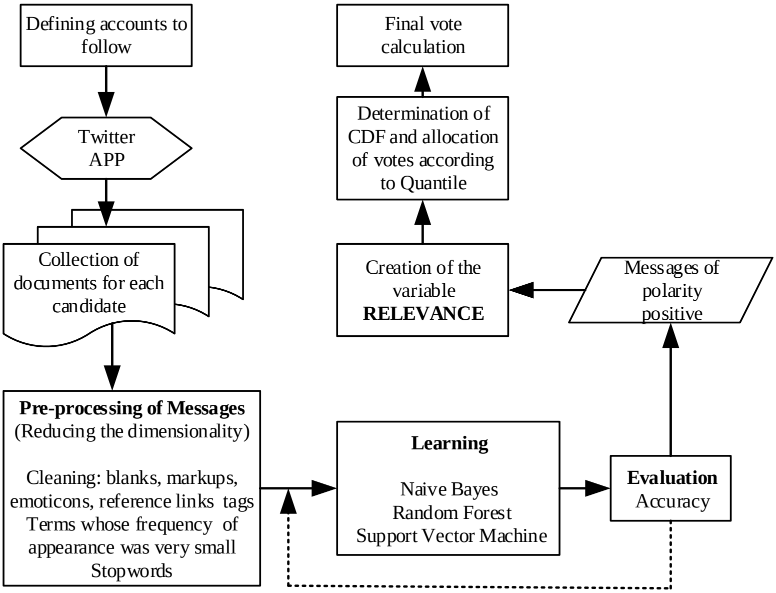
Defining accounts to follow
Twitter APP
Collection of documents for each candidate
Pre-processing of Messages
(Reducing the dimensionality)
Cleaning: blanks, markups, emoticons, reference links tags
Terms whose frequency of appearance was very small
Stopwords
Learning
Naive Bayes
Random Forest
Support Vector Machine
Evaluation
Accuracy
Messages of polarity positive
Creation of the variable RELEVANCE
Determination of CDF and allocation of votes according to Quantile
Final vote calculation

Figure 2. Diagram of the methodology used. Source: Authors (2018).

3.5. Learning

Supervised classifiers require that the dataset be trained and then tested. The training of the dataset was done using the criteria of an expert manually, in which a sample of tweets was selected and manually classified according to the subjective value of each message in two categories, for or against of each of the candidates.

The training data are clean vectors, formed by the words of a tweet, to which they have been trained manually with a polarity value; and that function is validated with a set of test data. Once the classifier has the desired precision, this function is used to classify the rest of the data.

Next, with the training data set, the respective test of each dataset was carried out using each of the methods considered: Naive Bayes, Random Forest and Support Vector Machine.

3.6. Evaluation

For the evaluation, the Accuracy metric was used, which in general terms accounts for the overall presition in the performance of the classification method.

3.7. Messages of polarity positive

Once the respective dataset of each candidate was evaluated, only those messages that had positive polarity were considered.

3.8. Creation of the Relevance variable

Relevance is an indicator that determines the degree of importance that an object may have, in terms of communication, it can be said that one message is more relevant than another because of the level and influence it may have on the voters. Within the opinion leaders this influence is due to two factors fundamentally, its popularity and its prestige.

Our proposal to measure relevance is given by a function that depends on the followers you have, the degree of acceptance and the dissemination of the message.

But what can affect the relevance of a message in the political work ?. Basically in the adepts and in the potential followers that this can generate.

The makers of opinion are not necessarily the mass media, nor the politicians themselves, nor the opinion leaders; but they are ordinary people. Take the example of a father of a family, who with his authority affects the political decision of his children; Take also the example of a message that has a high degree of relevance, the followers of it make it their own and relay it to another, becoming a channel of dissemination. It is logical to suppose then that to a greater degree of relevance of an adept and of a message this one must have a reward, that becomes to gain adepts of such message and therefore in votes by a political option.

We have quantified this vote gain, or conversion given by the relevance in potential votes that can be generated. Those messages that are more relevant will have a greater impact on the electorate than those that do not have as much relevance. This way of quantifying or converting the votes is assumed to apply to those voters who have not yet decided who to vote for, the undecided ones, who are usually in a large majority in electoral campaigns and as the elections approach they diminish to the extent that they are aligning themselves with one or another position. This way of going ascribing to one or another position is given for many reasons, the one that interests us is the one that is based on relevant information that these undecided have at their disposal and that affects them in such a way that makes them adopt a position defined.

For Twitter several metrics have been defined on the relevance of the messages and they are used to measure the importance of a Twitter account, the most used is the one that measures the number of followers or favoriteCount, but we also want to measure the diffusion of the political messages as a mechanism of persuasion in other voters. As Twitter is increasing the amount of information continuously, then it is necessary to measure this expansion using another metric, that of the messages that are forwarded, retweetCount. In the present investigation we have defined the Relevance in function of these described parameters, giving greater importance or weight to the forwarding of the messages, to the extent that they serve as diffusers of the message. In equation 1 this metric is defined.

$$Relevance = \frac{favoriteCount + 3retweetCount}{Ntweets} \qquad (1)$$

This metric that gives a higher score to those messages that have greater diffusion and that also come from accounts that have a large number of followers applies only to those tweets with positive polarity, which are in favor of a certain candidate.

3.9. Determination of CDF and allocation of votes according to Quantile

Once the calculation of the Relevance variable is determined, we evaluate its density function, given that this measure is a random variable, given that it comes from a biased sample of voters, made up of those citizens who send their messages on Twitter. The density function that best fits the data of the Relevance variable is a Johnson type Sl, whose parameters are: (Katz, 2009)

Table 1. Parameters of Johnson Sl Distribution of variable Relevance of messages.

Type	Parameter	Estimation
Shape	γ	1,143
Shape	δ	0,012
Localitation	θ	0,000
Scale	σ	1

Source: Authors (2018)

We used the Kolmogorov-Smirnov-Lilliefors test, K-S-L, (Pedrosa, 2015) for the goodness of fit and we accepted the null hypothesis H_0 = *the data come from a Johnson Sl distribution.*

With this density function, then, it's respective CDF cumulative distribution function was used to determine the thresholds for which we would give the conversion of the votes given by the value of the relevance. We use the quantile 97,5%, $Q_{97,5\%}$, as a threshold, over which if the Relevance is greater, that voter would be given a total of three votes, since his message could affect two other people, if the Relevance is greater than the 95% quantile, $Q_{95\%}$, but less than $Q_{97,5\%}$ will be assigned two votes, the one may affect one more voter. For values lower than the $Q_{95\%}$ we assumed that it would not have an impact on other voters. In Figure 3, the distribution of the Relevance variable and the vote allocation scheme are shown.

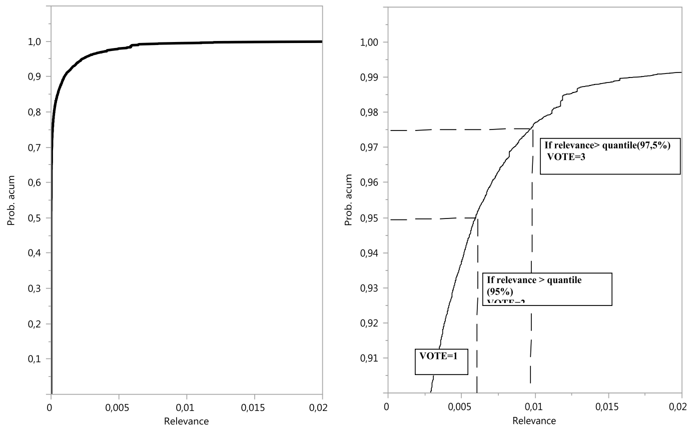

Figure 3. Johnson SI Distribution the variable Relevance and assignment of votes. Source: Authors (2018).

3.10. Final vote calculation

The final calculation of the votes is presented in the next section.

4. Results

Once for each Twitter message has been determined its polarity and its assessment in votes has proceeded to obtain the results; As shown in Table 2, the comparison between the official results given by the National Electoral Council of Ecuador, CNE, and those obtained using the defined Relevance methodology is presented. An MAE = 1.4 compared to the prediction made analyzing the Twitter messages with the proposed methodology using the valuation of votes given by the variable Relevance make this methodology an alternative in the electoral prediction.

Table 2. Comparison of results between Official results and the method using Relevance factor.

Candidate	Official Result CNE	Using the Relevance factor
Moreno	39,4	38,3
Lasso	28,1	29,7
Viteri	16,3	14,5
Moncayo	6,7	6,0
Others	9,5	11,5
MAE		1,4

5. Conclusions

The present study tries to demonstrate that the analysis of the tweets emitted by the users about their electoral preferences is as reliable as the results issued by different surveys.

The contribution of this research is the incorporation of the variable Relevance to enhance the electoral vote.

Additionally, the cost involved in the application of this methodology and the survey is incomparable; to more than the speed in the delivery of results.

References

Blasquez, D. & Domenech, J., (2017). Big Data sources and methods for social and economics analyses. *Technological Forecasting and Social Change*.

Ceron, A., Curini, L., & Iacus, S., (2015). Using social media to forecast electoral. *Statistica Applicata Italian Journal Of Applied Statistics*, 239-261.

Eco, U. (10 de 06 de 2015). *lastampa*.

Katz, E. & Lazarfeld, P., (2009). *Personal Influence, the Part Played by People in the Flow of Mass Communications.* New Jersey: Transaction Publishers.

Pedrosa, I. J.-B.-F.-C. (2015). Pruebas de bondad de ajuste en distribuciones simétricas. *Universitas Psychologica*, 245-254.

Singh, P., & Sawhney, R., (2018). *Progress in Advance Computing and Intelligent Engineering.* Singapore: Springer Nature Singapore.

Tumasjan, A. S. (2010). Predicting Elections with Twitter. What 140 Characters Reveal about Political Sentiment. *Fourth International AAAI Conference Weblogs and Social Media*, (págs. 178-185).