Twitter Sentiment Analysis applied to the 2017 Chilean Elections

Tomás Alegre Sepúlveda
Universidad Católica del Norte
Av. Angamos 0610, Antofagasta-Chile.
Email: tas001@alumnos.ucn.cl

Brian Keith Norambuena
Universidad Católica del Norte
Av. Angamos 0610, Antofagasta-Chile.
Email: brian.keith@ucn.cl

ABSTRACT

Every four years, around 14 million Chileans are called to exercise the right to vote in presidential elections. Due to the fact that in recent years several candidates run for the presidency and social networks are a very important source of information in the current times, in this work we study the process of sentiment analysis, through the behavior shown in Twitter generated in the 2017 election period in Chile, based on three candidates, Sebastián Piñera, Beatriz Sánchez and Alejandro Guillier. This analysis has been carried out with text mining and classic classification methods, such as Naïve Bayes and Support Vector Machines. The obtained results show that this approach can be successfully applied to obtain the positive or negative orientation of a tweet regarding the 2017 Chilean elections.

Keywords: Sentiment Analysis, Elections, Social Networks, Machine Learning.

INTRODUCTION

Nowadays, social networks are an important part of the digital life of different users, this leads to the possibility of finding different opinions on some activities within these networks. Specifically, Twitter is one of the most important and busiest social networks today, with nearly 400 million registered users in 190 countries [1]. Within Chilean society, there is a broad range of opinion on the candidates and their respective electoral processes, either for or against them. Social networks offer another perspective for the analysis of public opinion, using tools provided by social networks. This work uses the classification methods of Support Vector Machines (SVM) [2] and Naïve Bayes (NB) [2, 6] in order to predict the orientation (positive or negative) of a tweet regarding the 2017 Chilean elections. It should be noted that the work and the results shown within the research is an intermediate work and the final objective is the production of a way to estimate the voting intention using the tweets associated with each of the users belonging to the Twitter platform.

Related works

In the work described in [8] there is research on the classification of Tweets using hybrid methods combining SVM with CNN (Convolutional Neural Network) where the hybrid method gets the best results. In particular, the authors obtain an accuracy of 61.7 % with SVM, 64.1 % with CNN and 64.7 % with the hybrid method. In addition, it describes the difficulties with the classical classification methods at the time of classification of neutral tweets, which cause the accuracy to be negatively affected. In this context, the exclusion of neutral tweets in this work is justified.

In other research, comparisons are made between different methods, as in the case of [7]. This paper compares the methods of SVM and SentiStrength (dictionary-based classifier) and explains the advantages and disadvantages of each of these text classification methods. In short, you can see research on different articles and their percentage efficiency in terms of accuracy in classifying each of the texts. The case study was conducted mainly on the basis of data associated with comments made on Twitter and, in general, most of the results were close to 75% for accuracy for both classifiers.

In [9] different problems are exposed about carrying out research on Twitter, together with different works by other authors who carry out research on the social networks mentioned above, where there is talk about the biases that exist in the population that uses Twitter and that this may not be representative or that does not adapt to reality. Along with this, the author recommends using more robust classifiers for this type of research, since Twitter users tend to use a lot of humor in their texts, which makes it difficult to use classic classifiers.

Data extraction

The data was extracted from Twitter, using its Application Programming Interface (API) together with the JavaScript programming language. Twitter is one of the most important microblogging platforms on the internet, within this platform each user can update their profiles in 280 characters. Due to this limitation, the use of hashtags has gained popularity, being considered a very important feature in many tweets. This simple way of using the platform contributes to its success, as users can write about their lives, give opinions and discuss any topic that comes to their mind.

Specifically, the API was used to Search Tweets on the Full Archive, which applies filters such as time period and hashtags. With this tool, it is possible to search for indexed tweets from 2006 onwards with a total of 5,000 tweets per period [3].

The data extraction process was limited by Twitter standards on its API and the restrictions on the requests that could be generated. *Cronjobs* were used in the system to execute the script from time to time with pre-established query parameters without the need for human intervention.

Tweets were extracted for the candidates Sebastián Piñera, Beatriz Sánchez and Alejandro Guillier and the time periods were from September 1st, 2017 to October 31st, 2017 dividing the tweets into a 60% training and 40% test sets.

Manual classification of tweets

Within the work of manual classification of each of the tweets must take into consideration some important elements, mainly the classification is made based on keywords within the text associated with the tweet, among them it is possible to find specific words associated with some positive or negative emotion. Observing the context in which these words are found it is possible to discern the polarity of the text, on the other hand, there is within the idioms of Twitter what is called hashtags, which contain concatenated text in a way that expresses a stronger emotion and facilitate the detection of the polarity of the text for the person who is doing the task of classifying tweets.

Preprocessing of the Data

In sentiment analysis it is usually necessary to normalize the text data before applying analysis techniques. In general, this is one of the main tasks to be performed for natural language processing [4]. The text normalization can be done by extracting punctuation marks, converting words to lowercase, eliminating special characters and removing words that do not provide any semantic information, also known as stopwords [5]. After eliminating each one of the elements described above, the canonical form of the texts is obtained.

In the collected data, a raw text input was found, which contains several elements that do not help the analysis such as punctuation, hashtags, and others. Thus, the following preprocessing were considered: the text format is decoded from UTF-8; all URLs are removed; words with accents and special characters are modified to their canonical form; all the digits of the text are deleted; the characters of the text are converted to lowercase; special characters are removed (/#"$;._-); users are removed from the text; repeated letters are eliminated from the text leaving only 2 letters (e.g., hoooola - hoola); exclamation and question marks are eliminated; empty words and stopwords are eliminated, and stemming is applied to the words of the tweet.

RESULTS AND DISCUSSION

Two methods, Naïve Bayes and SVM, were applied to the three candidates. The results will be shown for each one of these candidates separately and for all candidate together (overall results). The results were also divided into 15-day periods from September 2017 to October 2017.

Overall results

Initially, a total of 9,600 tweets were extracted, which went through several stages of filtering (eliminating repeated tweets, empty tweets, and irrelevant tweets) and preprocessing as described previously. In general, the number of total tweets after this filtering and preprocessing was 947, the Naïve Bayes and SVM methods were applied to these. The results obtained can be seen in Table 1 for the corresponding test set (30% of the total tweets).

Table 1 Overall results evaluation metrics for the test set.

Clasiffier	Accuracy	Recall	Precision	F1
SVM	76.45%	76.41%	76.89%	76.65%
NB	66.31%	77.99%	62.82%	69.59%

Candidate Sebastián Piñera

3,200 tweets were collected for the candidate Sebastián Piñera (SP), and after cleaning the text, 239 tweets were obtained and divided into 15-day periods. For SVM, the results can be seen in Figure 1, together with Table 2. For Naïve Bayes the results can be seen in Figure 2 together with Table 3.

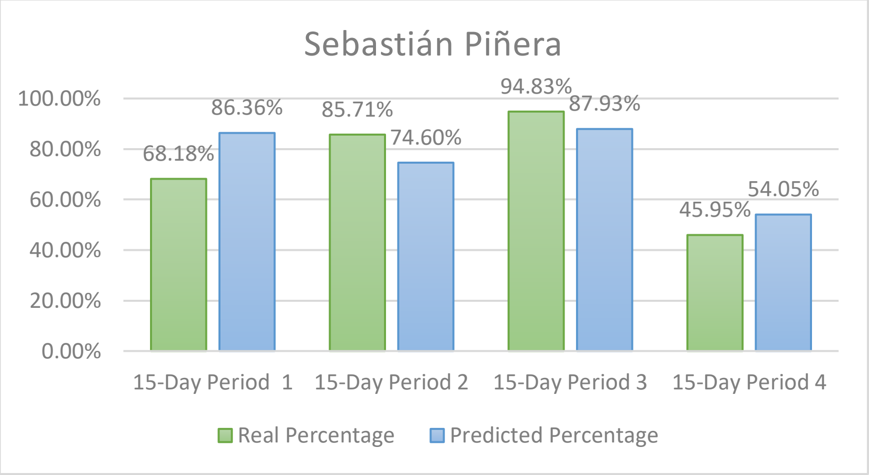

Figure 1 Real percentage vs predicted percentage of positive tweets in every 15-day period for SP with SVM.

. Table 2 Results evaluation metrics for SP with SVM.

SP	Accuracy	Recall	Precision	F1
15-Day period 1	72.73%	93.33%	73.68%	82.35%
15-Day period 2	82.54%	83.33%	95.74%	89.11%
15-Day period 3	89.66%	90.91%	98.04%	94.34%
15-Day period 4	67.57%	73.53%	62.50%	67.57%

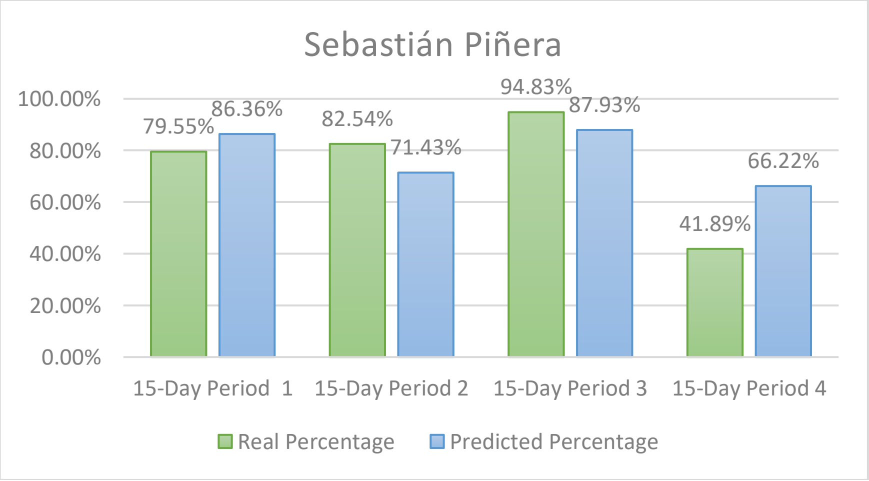

Figure 2 Real percentage vs predicted percentage of positive tweets in every 15-day period for SP with NB.

Table 3 Results evaluation metrics for SP with NB.

SP	Accuracy	Recall	Precision	F1
15-Day period 1	75.00%	88.57%	81.58%	84.93%
15-Day period 2	73.02%	76.92%	88.89%	82.47%
15-Day period 3	86.21%	89.09%	96.08%	92.45%
15-Day period 4	56.76%	77.42%	48.98%	60.00%

Candidate Beatriz Sánchez

For the candidate Beatriz Sánchez (BS), 3,200 tweets were collected and after cleaning the text, 407 tweets were obtained divided into 15-day periods. For SVM, the results can be seen in Figure 3 together with Table 4, while for Naïve Bayes, the results can be seen in Figure 4 together with Table 5.

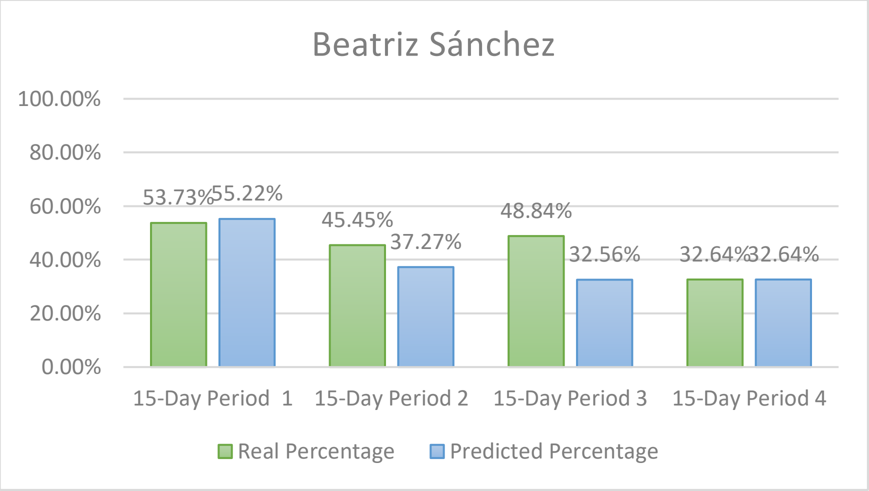

Figure 3 Real percentage vs predicted percentage of positive tweets in each 15-day period for BS with SVM.

Table 4 Results evaluation metrics for BS with SVM.

BS	Accuracy	Recall	Precision	F1
15-Day period 1	80.60%	83.33%	81.08%	82.19%
15-Day period 2	73.64%	62.00%	75.61%	68.13%
15-Day period 3	72.09%	54.76%	82.14%	65.71%
15-Day period 4	79.17%	68.09%	68.09%	68.09%

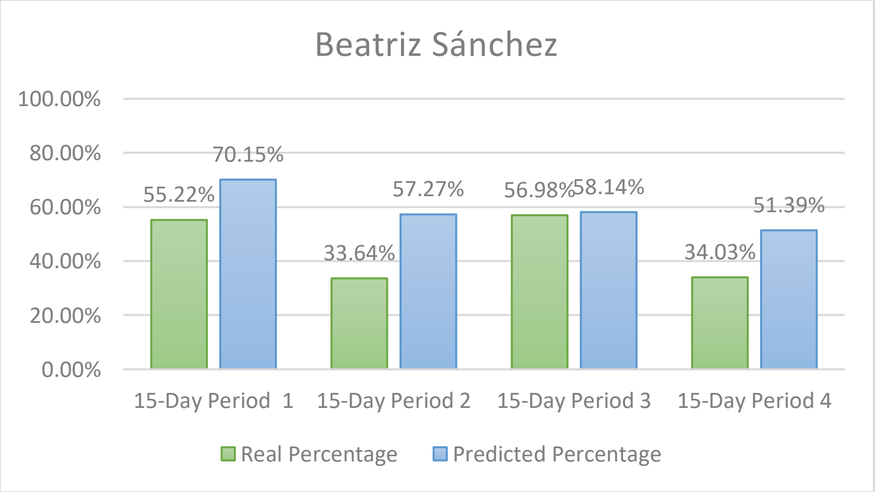

Figure 4 Real percentage vs predicted percentage of positive tweets in each 15-day period for BS with NB.

Table 5 Results evaluation metrics for BS with NB.

BS	Accuracy	Recall	Precision	F1
15-Day period 1	61.19%	78.38%	61.70%	69.05%
15-Day period 2	65.45%	83.78%	49.21%	62.00%
15-Day period 3	70.93%	75.51%	74.00%	74.75%
15-Day period 4	65.97%	75.51%	50.00%	60.16%

Candidate Alejandro Guillier

For the candidate Alejandro Guillier (AG) 3,200 tweets were collected, and after cleaning the text, 301 tweets were obtained divided into 15-day periods. The results for SVM can be seen in Figure 4 together with Table 6. While for Naïve Bayes the results can be seen in Figure 6 and Table 7.

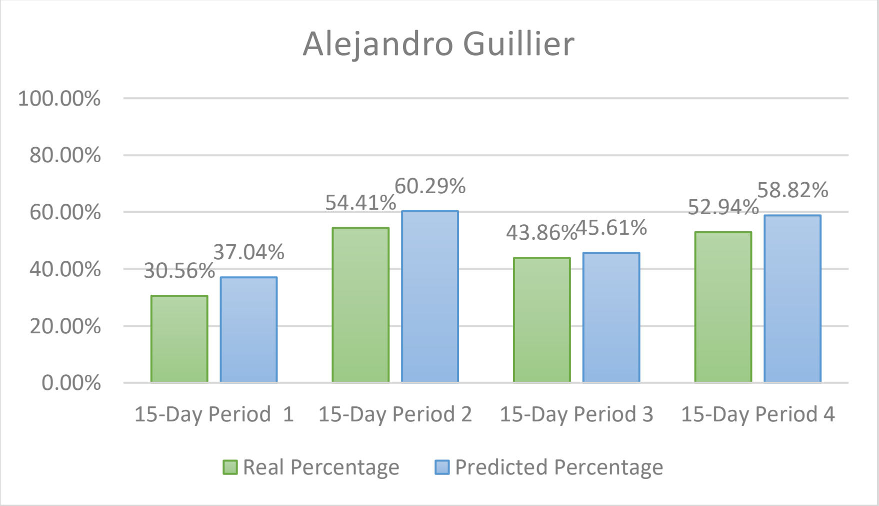

Figure 5 Real percentage vs predicted percentage of positive tweets in each 15-day period for AG with SVM.

Table 6 Results evaluation metrics for AG with SVM.

AG	Accuracy	Recall	Precision	F1
15-Day period 1	75.00%	69.70%	57.50%	63.01%
15-Day period 2	79.41%	86.49%	78.05%	82.05%
15-Day period 3	70.18%	68.00%	65.38%	66.67%
15-Day period 4	76.47%	83.33%	75.00%	78.95%

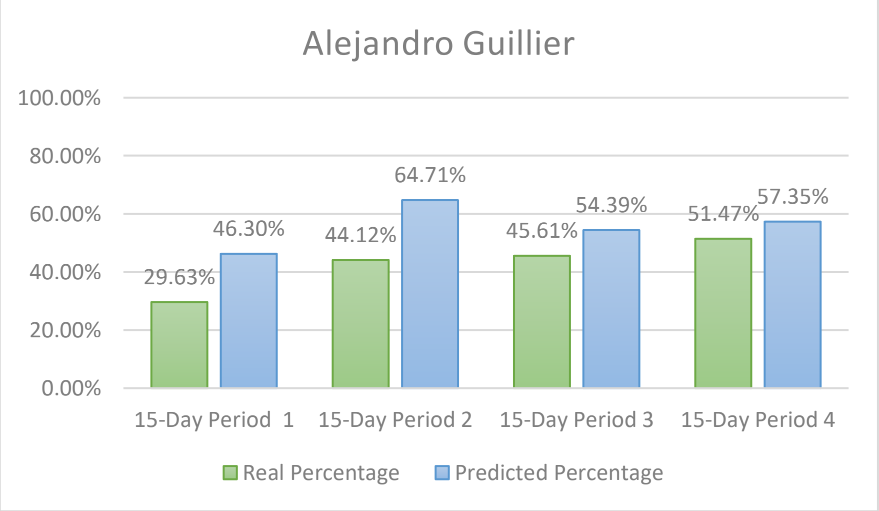

Figure 6 Real percentage vs predicted percentage of positive tweets in each 15-day period for AG with NB.

Table 7 Results evaluation metrics for AG with NB.

AG	Accuracy	Recall	Precision	F1
15-Day period 1	59.26%	59.38%	38.00%	46.34%
15-Day period 2	52.94%	70.00%	47.73%	56.76%
15-Day period 3	73.68%	80.77%	67.74%	73.68%
15-Day period 4	67.65%	74.29%	66.67%	70.27%

Discussion of results

Several interesting facts on the elections can be noted. First, between the two methods that were used, the SVM method shows better results in terms of prediction accuracy for the orientation of the tweets.

Regarding the candidates, it can be observed that the results for Beatriz Sánchez tend to a percentage decrease of positive tweets in each of the time periods, which could indicate that her popular support decreased enough to see the result reflected in the elections.

On the other hand, the number of tweets of the candidate Sebastián Piñera is quite low compared to the other candidates, this is due to the various filters applied to each tweet, where a considerable amount was eliminated due to several factors, such as the fact of not having useful content for the investigation (e.g., repeated tweets).

An observation of interest is the difference between the percentage of positive tweets between Alejandro Guillier and Sebastián Piñera, in particular, in the last period of evaluation (fourth 15-day period), there are similar results, which would correspond intuitively with the existence of a tie in terms of public opinion.

Finally, comparing the percentage graphs of Sebastián Piñera and Alejandro Guillier, it can be observed that regarding the number of positive tweets of each candidate there is an inverse relationship between periods until reaching a similar point in the last period (ie, the support on Twitter seems to alternate between the candidates every 15-day period, until becoming approximately equal in the last one).

CONCLUSIONS

This work has presented the application of sentiment analysis on the 2017 Chilean elections using classical machine learning classification. From the analysis of results, it has been found that it is possible to use social networks and sentiment analysis in order to notice social trends. It is also possible to emphasize that, using the Python programming language and the various associated libraries, it is possible to build a working sentiment classifier in a reasonable amount of time. Through this, it is possible to gauge public opinion according to the tweets for presidential candidates using text mining.

Regarding the data analyzed, it should be noted that the period with the most interaction on Twitter was from October 15 to October 31, corresponding to the last period analyzed (i.e., the period closest to the first-round election date).

With respect to the results obtained, it should be noted first that the SVM classification method presented better results in comparison with the Naïve Bayes method. While, for candidates, it should be noted that the number of tweets of Beatriz Sánchez is significantly higher than the other candidates for the presidency, on the other hand, the number of tweets of Sebastián Piñera is significantly lower than the rest of candidates, since several exact repetitions (or other similar anomalies) were found that were filtered in each case.

On the other hand, there are different expectations related to future research work, first of all, the implementation of different classifiers to improve classification performance is considered. However, the main line of future work is to investigate the possibility of generating a conversion for positive tweets to the general voting intention for the different candidates.

In addition, in the work carried out it was possible to find certain important problems, among them the main one was the existing limitations within the Twitter platform, the tool of this social network only allows to make requests for 5000 tweets on a monthly basis. This reduced the potential number of tweets available for analysis.

Finally, it should be noted that there is a possibility of biases in the tweet labeling process because the textual classification depends on the person in charge of this task and their psychological aspects and political opinion. For this reason, it is necessary to take these limitations into account when trying to generalize or apply these results.

REFERENCES

[1] Semiocast, «Geolocation of Twitter users,» Semiocast, 2012. [En línea]. Available: https://semiocast.com/en/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US. [Último acceso: Julio 2018].

[2] Twitter, «Twitter Developer Platform - Twitter Developers,» Twitter, [En línea]. Available: https://developer.twitter.com/content/developer-twitter/en.html. [Último acceso: 2018].

[3] Liu y Bing, Web data mining: exploring hyperlinks, contents, and usage data, Springer Science & Business Media, 2011.

[4] Leskovec, Jure and Rajaraman, Anand and Ullman y Jeffrey David, Mining of massive datasets, Cambridge university press, 2014.

[5] Sanderson, Mark and Christopher, D and Manning y Hinrich and others, «Introduction to information retrieval,» *Natural Language Engineering,* vol. 16, nº 1, p. 100, 2010.

[6] Paltoglou, Georgios and Thelwall y Mike, «Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media,» *ACM Transactions on Intelligent Systems and Technology (TIST),* vol. 3, nº 4, p. 66, 2012.

[7] T. Baviera, «Técnicas para el Análisis de Sentimiento en Twitter: Aprendizaje Automático Supervisado y SentiStrength,» *Revista Dígitos,* vol. 1, nº 3, pp. 33-55, 2017.

[8] A. Rosá, L. Chiruzzo, M. Etcheverry, and S. Castro, "Retuyt in tass 2017: Sentiment analysis for spanish tweets using svm and cnn," arXiv preprintarXiv:1710.06393, 2017.

[9] D. Gayo-Avello, «"I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper"--A Balanced Survey on Election Prediction using Twitter Data,» *arXiv preprint arXiv:1204.6441,* 2012.