







Sentiment Analysis in the Ecuadorian Presidential Election

Juan Carlos Minango Negrete¹(✉) , Yuzo Iano² ,
Pablo David Minango Negrete², Gabriel Caumo Vaz² ,
and Gabriel Gomes de Oliveira² 

¹ Instituto Tecnológico Universitario Rumiñahui, Sangolquí, Ecuador
juancarlos.minango@ister.edu.ec

² State University of Campinas, Campinas, Brazil
yuzo@unicamp.br, oliveiragomesgabriel@ieee.org

Abstract. The social network Twitter is characterized as one of the main means of communication used by politicians, governmental agencies, among others to communicate and engage with the people and prospective voters. This article proposes a methodology that might be employed in text extraction, acquisition, cleaning, and processing with the aim of analyzing the sentiments in the texts within the tweets of the 2021 Ecuadorian presidency candidates, Andrés Arauz and Guillermo Lasso. Many libraries that are specialized in text analysis were used to determine the sentiment analysis in the extracted tweets. The achieved results show a classification comparison among the employed libraries, highlighting that most of the tweets of both candidates were classified as positive and, following, neutral polarity.

Keywords: Ecuador · Election · Data analysis · Twitter

1 Introduction

With the emergence of the so-called Web 2.0, the massive production of text-type information on web pages arose, and it was boosted since the advent of social networks, which changed the way people communicate and interact with each other [1].

Companies, from the smallest to the biggest ones, have been making every effort and investment to create multimedia content through posts, comments, and messages, which allow the users to interact and meet such companies. Similarly, public people, politicians, and artists, among others, use social networks to share their opinions, sentiments, and even their political position.

Under this context, Twitter is a microblogging social network that allows 280 characters per tweet and has about 187 million active users per month [2]. These characteristics make Twitter the ideal platform for users to share their opinions

about worldwide tendencies. Therefore, Twitter is mainly used by government agencies and politicians, among others, as an environment to share news, ideas, and opinions.

Twitter is a trustworthy source to perform sentiment analysis [3]. Sentiment analysis is a computational process that has the task of processing, identifying, and classifying the opinions expressed by someone in a text to determine the person's position concerning a specific product, event, etc. Three sorts of sentiments are considered: positive, negative, and neutral [4].

This case study assesses people's stance with respect to the second round of the Ecuadorian presidential election in 2021, which was disputed by candidate Andrés Arauz and his opponent Guillermo Lasso. The data collection was made during the electoral campaign period.

The article is organized as follows: Sect. 2 makes an overview of works related to sentiment analysis. In Sect. 3, the applied methodology, the data collection and pre-processing, and the libraries used to perform the sentiment analysis are described. Section 4 groups the results achieved with a comparison of the libraries employed in the sentiment analysis. Finally, Sect. 5 presents the conclusions.

2 Related Works

In [5], the tool Stanford NLP for Natural Language Processing (NLP) was applied with an adaptation for Ecuadorian regional language. The sentiment analysis was performed on the social network Twitter with the messages sent to Ecuadorian former president Rafael Correa. The political assessment via Twitter in South America is still very scarce, and the election analyses in Venezuela [6], Chile [7], and Argentina [8].

Works about sentiment analysis based on the English language, like [9], study the governmental policies from the point of view of ordinary people in order to analyze their sentiments and stances concerning the new demonetization rules. The data collection is made via trends query using hashtags (`#demonetization`).

Finally, it is worth highlighting that most sentiment analysis libraries were developed for the English language due to its universality.

3 Methodology

This section presents the proposed work methodology to analyze and extract information from the social network Twitter data source. So, Fig. 1 shows the workflow applied in this experimental research.

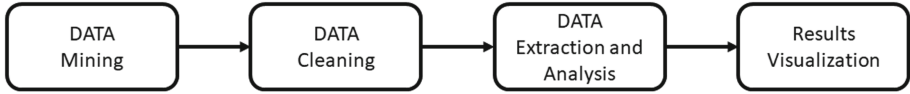


Fig. 1. Work methodology.

3.1 Data Mining

Depending on the software employed in the data collection from Twitter, there are many libraries and APIs of the own social network, which allow the extraction of countless characteristics embedded in each tweet of the user.

To use the Twitter proprietary API, it is necessary to create a register, specifying the reasons and the aim of the use of the data. Once the register is made, it must be approved to make the *Keys* and *Tokens* available and allow the exploration or mining of the information.

In this work, it was employed the development interface of Python with the IDE Jupyter Notebook, which has the library *tweepy*. The data extraction was directed to the search of current tendencies in the political sphere: hashtags that had trends in the Ecuadorian election campaign were evaluated, specifically for the second electoral round, which took place from March 6th, 2021 to April 8th, 2021.

The library *tweepy* provided all the information embedded in each tweet, which was subsequently stored in a CSV (Comma Separated Values) file within which the following fields were chosen for the analysis:

- Tweet text - “text”.
- Tweet creation date - “created_at”.
- Tweet location - “location”
- Retweets amount - “retweet_retweet_count”

3.2 Data Cleaning

Once the tweets that refer to the searched trend were achieved, the text of each tweet was cleaned, and the variables formats were adjusted to proceed with the right analysis in each of them. The CSV file was loaded in the format *data frame* with the help of the library *pandas*. To remove duplicate tweets, it was applied the tool `drop_duplicates`, which perform a row-by-row analysis in the specified column aiming to delete the duplicate rows.

Continuing with the data pre-processing, a function was created to clean the text based on regular expressions. Furthermore, it is considered the removal of links for external web pages with format `http\S+`, which allows finding this format within the text and delete the whole string that follows it. In the regular expressions, punctuation marks that affect the correct understanding of the context of the tweets (e.g., periods, commas, question marks, and exclamation marks, among others) were considered.

In addition, the removal of users named within the tweets was also considered. To perform such elimination, regular expressions including the format `@\S+` were used, allowing the deletion of user names mentioned within the text of the tweet.

StopWords: StopWords are lists of words stored in the package NLTK for the natural language text pre-processing, which aim to remove a set of words that do not affect the phrase context. The StopWords are, mainly, articles and pronouns that, when removed, do not affect the phrase context in an afterward analysis.

Tokenization: This is another NLTK package that, once all the aforementioned pre-processing is performed, breaks the phrases in each one of the words that compose them. The tokenization allows the computer to analyze the text word by word.

3.3 Frequency of Words - Word Clouds

After the data cleaning, the NLTK tool to compute the frequency of the words was applied. This tool counts the words that repeat more frequently within the texts of all tokenized tweets.

In contrast, the word cloud graphic shows visually the words' frequency. It presents the more frequent words with larger font size, i.e., the higher the number of mentions within the analyzed text, the larger the font size in the word cloud graphic.

3.4 Sentiment Analysis in Spanish

For the sentiment analysis in the Spanish language, the library *sentiment-analysis-spanish* (version 0.0.25) was used. It is a library that applies a Naive Bayes classifier to predict the sentiments in the Spanish phrases. It was trained with 800,000 comments from users of the platforms Decathlon, Tripadvisor, FilmAffinity, ElTenedor, and eBay, which were extracted through web scraping. This library achieved an accuracy of 90% on the test dataset.

Equation 1 represents the Naive Bayes classifier, which is given by the conditional probability, where c is the text class (i.e., the kind of sentiment, which may be positive or negative), and x is the tokenized text (each word that composes the phrase have a numerical value that makes possible the computation of the probability of belonging or not to a specific sentiment, which might be positive or negative).

$$P(c | x) = \frac{P(x | c) P(c)}{P(x)} \quad (1)$$

This library computes the probabilities of the assessed phrase with values from 0 to 1. When the phrase or text achieves values that are close to 1, it is classified as positive. Otherwise, low probabilities are classified as negative. In this case study, values within the range from 0.4 to 0.6 are considered neutral.

3.5 Sentiment Analysis with English Translation

Since the sentiment analysis libraries for the English language are more developed, the tweets extracted in Spanish were translated to English.

For this purpose, the library *TextBlob*, which uses an API of Google Translate to translate texts to different languages, was employed. There is a daily limit for the number of translations, so the set of tweets was split in order to be possible to translate the texts without reaching the daily amount of translations allowed by the API.

TextBlob is a library of natural language processing that includes a class of sentiment analysis to determine the text polarity and its subjectivity as well. For the text polarity, the library provides values in the range from -1 to 1 . When the values are close to -1 , the text is classified as negative. Otherwise, for values close to 1 , the classification is positive. In this study case, values that are close or equal to 0 are considered neutral.

Besides the library *TextBlob*, we also employed the library *Vader*, which was created specifically for the analysis of sentiments expressed in social networks, where the text syntax must be totally processed since there are many *posts* where the users use emoticons, colloquial language, and specific expressions from different regions. For that reason, the library *Vader* is based on the phrases lexicon and the manual construction of dictionaries of phrases and words that may help the text classification.

The library *Vader* delivers four values that represent the probability of a given phrase belong to a specific sentiment, which might be positive, negative, or neutral.

4 Results and Discussion

To analyze sentiments via the social network Twitter, it was employed, as the subject area, the Ecuadorian scenario during the second round of the 2021 presidential election, when it was possible to observe a growing use of social networks on the part of the presidential candidates Andrés Arauz from movement UNES (Unión por la Esperanza) and Guillermo Lasso from movement CREO (Creando Oportunidades).

The social networks (e.g., Twitter, Facebook, and the novel application TikTok) were the platforms where the candidates presented their proposals and attacked each other based on their prior history.

In this study case, tweets of both candidates were collected based on the social network trends and their hashtags. For the data mining of the candidate Guillermo Lasso, it was used the trend *#EncontremosParaLograrlo*, and about 2900 tweets that referenced this slogan were achieved. For the candidate Andrés Arauz, the same procedure was made with the trend *#ContigoConTodosAhora*, and about 1900 tweets were collected.

The first step of the data cleaning was the deletion of duplicated tweets. Afterward, the text pre-processing was performed with the NLP tools mentioned

in Sect. 3.2. Once the stop words and regular expressions were removed, a list with tokenized words was created to figure out the frequency of the words in the databases of both candidates.

Figure 2 shows the frequency of words for the trend #EncontremosParaLograrlo of the candidate Guillermo Lasso. In this figure, it is possible to distinguish that this hashtag has the greatest predominance in the tweets. Besides, one can note that the tweets tendency makes reference to the words “gobierno” (government), “progreso” (progress), “seguridad” (safety), “empleo” (employment), and others that are usually used in presidential proposals.

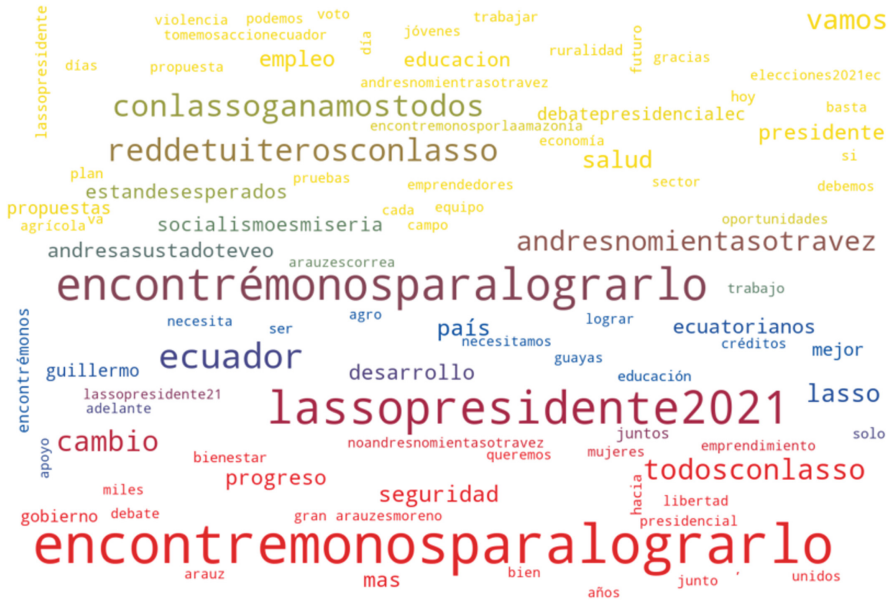


Fig. 2. Word cloud for #EncontremosParaLograrlo of candidate Guillermo Lasso.

Figure 3 presents the distribution graphics of the three methods employed in the classification of sentiments embedded within the tweets. It is worth highlighting that the library for sentiment analysis in Spanish *spanish_sentiment* (see Fig. 3 (a)) classified most of the tweets as negative or neutral. In contrast, the library *TextBlob* (see Fig. 3 (b)) classified many tweets as positive, followed by tweets with neutral polarity. In the case of the library *Vader* (see Fig. 3 (c)), most of the tweets were classified as positive.

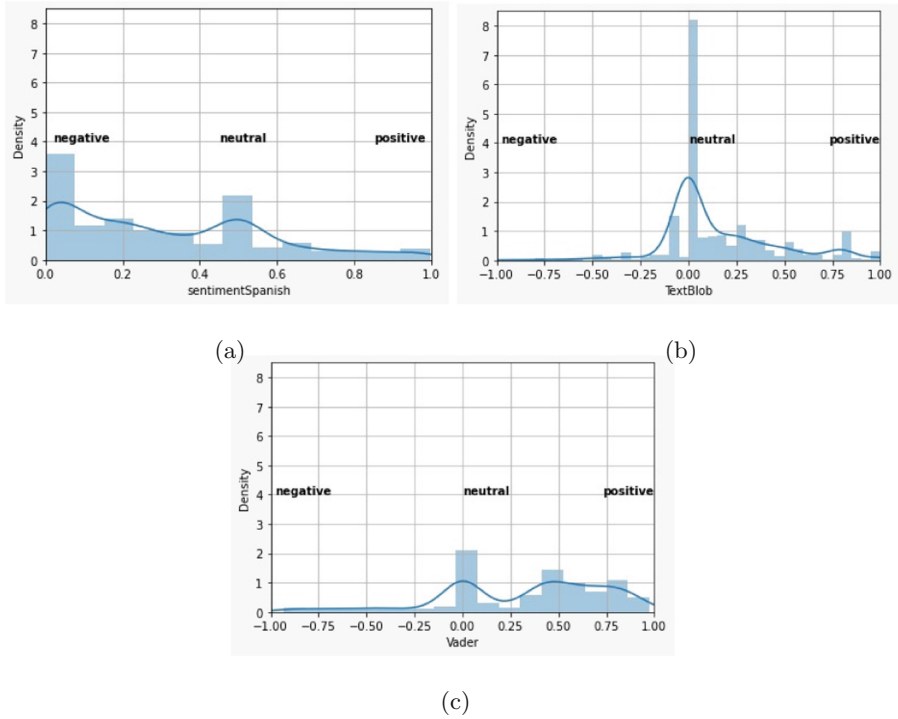


Fig. 3. Sentiment analysis in tweets collected for the trend *#EncontremosParaLorarlo* of candidate Guillermo Lasso. (a) *sentiment_spanish*, (b) *TextBlob*, y (c) *Vader*.

In Fig. 4, one can see a comparison of the three libraries employed in the sentiment analysis performed in the tweets collected from candidate Lasso. It is possible to observe that the libraries *TextBlob* and *Vader* effectively classify most of the tweets with positive polarity.

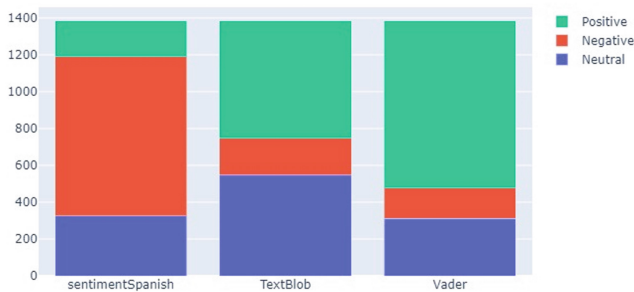


Fig. 4. Sentiment analysis with *spanish_sentiment*, *TextBlob*, and *Vader* of the candidate Guillermo Lasso.

Figure 5 shows the cloud of words with higher frequency in the tweets associated with the presidential candidate Andrés Arauz. One can observe that the most predominant word was not the hashtag trend, but the word “trabajo” (work). After analyzing the tweets that contained such a word, we could note that many of them make reference to the hard work that had been held for 20 months to strengthen the movement and reach the electoral second round since the candidate Andrés Arauz was totally unknown and took part for the first time in an election.



Fig. 5. Word cloud for #ContigoConTodosAhora of candidate Andrés Arauz.

In Fig. 6, there is a noticeable tendency of the libraries *TextBlob* (see Fig. 6 (b)) and *Vader* (see Fig. 6 (c)) to classify the tweets in the range from 0 to 1, siendo en su mayoría clasificados como neutros y positivos. i.e., most of them are classified as neutral or positive. In contrast, the library *sentiment_spanish* (see Fig. 6 (a)) classified most of the tweets as negative, followed by neutral.

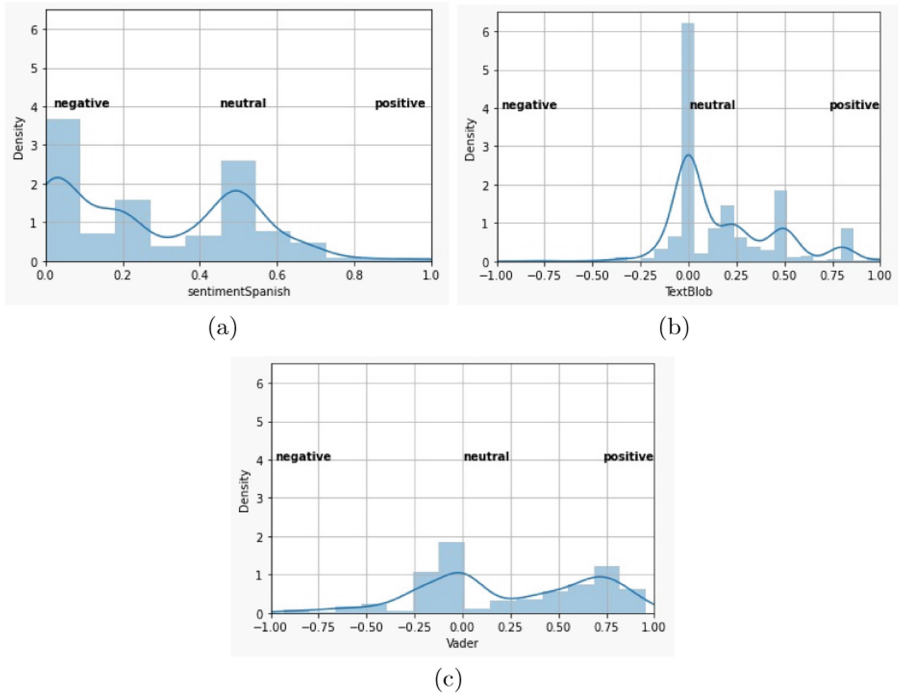


Fig. 6. Sentiment analysis in tweets collected for the trend `#ContigoConTodosAhora` of candidate Andrés Arauz. (a) `sentimentSpanish`, (b) `TextBlob`, y (c) `Vader`.

Figure 7 shows an overview of the way how each library of sentiment analysis classified the tweets of the candidate Andrés Arauz, where the tendency of the classifier `spanish_sentiment` was negative. By contrast, the libraries `TextBlob` and `Vader` classified most of the tweets with positive polarity.



Fig. 7. Sentiment analysis with `spanish_sentiment`, `TextBlob`, and `Vader` of candidate Andrés Arauz.

5 Conclusion

This article presented the process of data collection from the social network Twitter, the data cleaning, the text pre-processing, and the sentiment analysis with the aid of the libraries *sentiment_spanish*, *TextBlob*, and *Vader*. The data collection from Twitter had as scenario the presidential election that took place in Ecuador. The sentiment analysis performed with the library *sentiment_spanish* classified most of the tweets of both candidates as negative. In contrast, the libraries *TextBlob* and *Vader*, which have a better performance in the task of analyzing sentiments based on text, classified most of the tweets of both candidates as positive, followed by neutral polarity.

References

1. Asghar, M.: Detection and scoring of internet slangs for sentiment analysis using sentiWordNet. *Life Sci. J.* **11**, 66–72 (2014). <https://doi.org/10.6084/M9.FIGSHARE.1609621>
2. Statista Twitter: most users by country – Statista. Statista (2021)
3. Kouloumpis, E., Wilson, T., Moore, J.: Twitter sentiment analysis: the good the bad and the OMG!. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 5, no. 1, pp. 538–541 (2011). <https://ojs.aaai.org/index.php/ICWSM/article/view/14185>
4. Zahoor, S., Rohilla, R.: Twitter sentiment analysis using lexical or rule based approach: a case study. In: 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pp. 537–542 (2020). <https://doi.org/10.1109/ICRITO48877.2020.9197910>
5. Hidalgo, O., Jaimés, R., Gomez, E., Luján-Mora, S.: Sentiment analysis applied to the popularity level of the ecuadorian political leader rafael correa. In: 2017 International Conference on Information Systems and Computer Science (INCISCOS), pp. 340–346 (2017). <https://doi.org/10.1109/INCISCOS.2017.64>
6. Avila, D.A., Muñoz, A., Luengo, F., Chourio, X., Fernández, A.: Caracterizando las Elecciones Venezolanas a Través de Twitter. Caso: #26S. *Anuario Electrónico de Estudios en Comunicación Social. Disertaciones* **5**(1), 57–76 (2012). <https://www.redalyc.org/articulo.oa?id=511555573008>
7. Guevara, M., Pino, D., Mendoza, M., Silva, C., Olivares, M.: Chile y el Ecosistema de las Elecciones Políticas en Twitter (2013). <https://doi.org/10.13140/2.1.3324.7684>
8. Robins, D., Frati, F.E., Alvarez, J., Texier, J., Loto, L.: Balotaje Argentina 2015 a partir de un análisis de sentimiento de tweets. Zenodo (2016). <https://doi.org/10.5281/zenodo.51496>
9. Prabhsimran, S., Ravinder, S.S., Karanjeet, S.K.: Sentiment analysis of demonetization of 500 & 1000 rupee banknotes by Indian government. *ICT Express* **4**(3), 124–129 (2018). <https://doi.org/10.1016/j.ict.2017.03.001>