Predicting an election's outcome using sentiment analysis

Ricardo Martins, José João Almeida, Pedro Henriques, Paulo Novais
Algoritmi Centre / Department of Informatics
University of Minho, Braga - Portugal
ricardo.martins@algoritmi.uminho.pt, {jj, prh, pjon}@di.uminho.pt

Keywords: Natural Language Processing, Sentiment Analysis, Machine Learning

Abstract: Political debate - in its essence - carries a robust emotional charge, and social media have become a vast arena for voters to disseminate and discuss the ideas proposed by candidates. The Brazilian presidential elections of 2018 were marked by a high level of polarization, making the discussion of the candidates' ideas an ideological battlefield, full of accusations and verbal aggression, creating an excellent source for sentiment analysis. In this paper, we analyze the emotions of the tweets posted about the presidential candidates of Brazil on Twitter, so that it was possible to identify the emotional profile of the adherents of each of the leading candidates, and thus to discern which emotions had the strongest effects upon the election results. Also, we created a model using sentiment analysis and machine learning, which predicted with a correlation of 0.90 the final result of the election.

1 Introduction

It is undeniable that social media have changed how people contact to each other, enabling them to maintain relationships that previously would be difficult to maintain for various reasons, such as distance, the passage of time, and misunderstandings.

Barack Obama used social media as the main platform of his presidential campaign in the United States in 2008, and since then, it is well established that social media created a strong influence on voters' decisions in that election and many others. Facebook and Twitter are now considered essential tools for any political campaign, along with other social media platforms that have been created since then.

The influence of social media increases when candidates with low funding levels and low levels of traditional media exposure try to defeat adversaries with more resources. Social media allow the candidates to post their political platforms as well as inflammatory posts against their opponents. In many cases, political candidates use social media mainly to carry provocative attacks against their opponents. Their followers, in turn, use social media to broadcast their reactions of anger and satisfaction in posts that can be shared thousands of times.

In electoral campaigns highly marked by their massive presence on social networks - as was the Brazilian presidential election in 2018 – by using Natural Language Processing (NLP) and Machine Learning (ML) it is possible to identify what voters think and feel about the candidates and their proposals for the various areas of the government. Thus, some interesting research question that arises are: "How do the emotions about each candidate influence the electoral decision?" and "Is it possible to predict the result of an election only by knowing what voters "feel" about a candidate?"

In this paper, we present an approach to predict the results of the election. As a case study, we collected messages from social media about the Brazilian presidential election of 2018, where we used Sentiment Analysis, NLP and ML to identify which emotions dominated the electorate and their correlations with the number of messages about the candidates and finally predict the results.

The remainder of the paper is as follows: Section 2 presents some studies in sentiment analysis in elections that inspired this work, while Section 3 describes the process of dataset creation for our tests. Section 4 presents our experiments, explains the steps used in the analysis and discusses the results obtained from a set of tests performed. The paper ends with the conclusion and suggestions for future work in Section 5.

2 Related work

The analysis of emotions on Twitter to explain elections is not a new approach. Several researchers have already done work in this area, each with different approaches and results. Wang [10] developed a system to analyze the tweets about presidential candidates in the 2012 U.S. election as expressed on Twitter. His approach analyzes the text's polarities (positive, neutral or negative), the volume of posts and the word most used. This is the same approach used by Heredia et al [3], that trained a Convolutional Neural Network using an annotated lexicon - Sentiment140 - to detect the text's polarities.

These approaches do not go deeper on the reasons of polarities, and thus, does not identify which sentiments influence the voters' decision, which is the objective of our work.

Tumasjan [9] has developed an analysis of tweets for the German elections which, similar to Wang's work, used the polarities of phrases to analyze the messages. However, unlike the previous work, Tumasjan's study relates the volume of messages to the final result of the election. This approach is highly influenced by financial questions (as richer the candidate is, more publicity about him can be posted in social media), and for this reason, we did not consider trustworthy. Bermingham [1] used this same approach, using the data from the Irish general election of 2011, but, different than Tumasjan, he has expanded the model for training by using polls as parameters for training the predictions, which has inspired our work in the training dataset creation.

While the existing works focused only on the aspect of the text's polarities, the work of Martins [5] inspired our decision to consider the basic emotions contained in the text as a factor of influence in the decision of the voter, and use it to predict results.

3 Dataset creation

Initially, to draw from the data a general idea of the Brazilian voter, it would be necessary to generalize the emotional profile of these voters, regardless of their region. The idea is that, according to the emotions contained in the texts, it would be possible to determine relevant information about the characteristics of the Brazilian voters. Thus, a pipeline was created for dataset creation, as presented in Fig. 1.

This pipeline begins with a collection of messages about the candidates. For this purpose, we collected tweets from 145 cities in Brazil, with each state being represented by at least its four largest cities, and delimiting a radius of 30 km for each city. The option for geolocated tweets is to avoid posts from countries different than Brazil, where the author probably would be not able to vote in Brazilian's election. To select what would be considered relevant or not, we defined that only the tweets containing the main candidates' names would be collected. Tweets containing two or more candidates' names were analyzed for all candidates mentioned in the text. Moreover, we considered relevant only tweets - not retweets. This decision was inspired by the necessity to avoid viral posts or the ones from digital influencers. In other words, we wanted to know the opinion from the message's author about a candidate, not the opinion of an author who the author likes.

For gathering the tweets, we developed a script in Python using the official API provided by Twitter, and collected all geolocated messages from the 145 cities mentioned earlier in the period from May 2018 to October 2018, which contained at least one of the following names in their texts: "Bolsonaro", "Ciro Gomes", "Marina Silva", "Alckmin", "Amoêdo", "Álvaro Dias", "Boulos", "Meirelles" and "Haddad", divided in two groups: first round and second round.

3.1 Out of scope

After an initial analysis of the messages collected, we decided not to handle the ambiguity in the texts during the analysis. The reason for choosing not to address this problem was justified by the tiny amount of messages that could lead to erroneous interpretations. Since it was a mandatory requirement for the messages to contain the name of at least one candidate, the nature of the Twitter messages - which limits each post by 280 characters - already considerably inhibited this type of problem.

Furthermore, this initial analysis showed that the existence of the candidate's name in the text made the context of the message as political and egalitarian in the emotional sense, as derogatory nicknames for the leading candidate candidates bring negative emotions. So, we avoided that exacerbated emotional expressions of some candidates affect others.

3.2 Lexicon expansion

When working with sentiment analysis, a common approach is to use a dictionary-based algorithm to identify the emotional words in texts. However, according to Feldman [2], "the main disadvantage of any dictionary-based algorithm is that the acquired lexicon is domain-independent and hence does not capture the specific peculiarities of any specific domain." Thus, it is essential to know some particularities about the domain which the texts represent, to avoid misunderstandings and enable analysts to make a better classification of the sentiments contained in the texts.

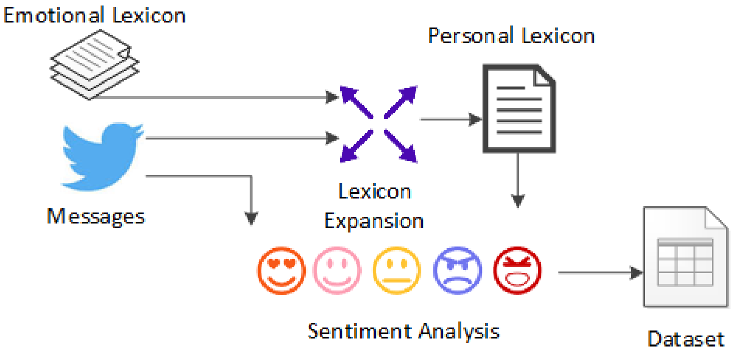

Figure 1: Dataset's creation pipeline

With this problem in mind, we adapted the solution presented by Martins [4], where the texts were represented by a vector of words and these vectors were used to analyze the similarities of the words contained in an emotional lexicon, to expand it. A major concern when creating these vectors was about the polarization among the candidates. Our idea was that the texts about a candidate do not influence the emotional words of other candidates. For this reason, we adopted the strategy of creating a personal emotional lexicon for each candidate and thus analyzing the candidate's sentiments individually according to their respective emotional lexicon.

An overview of the entire process of lexicon expansion is presented in Fig. 2.

3.2.1 Word Vectors

The process of lexicon expansion begins with grouping all tweets collected by the candidate's name, removing their stopwords and creating the word vectors. For this purpose, we developed a script in Python using the Word2Vec algorithm, presented by Mikolov [6], having as parameters: size of 50; window 5 and trained for 200 epochs. Later, the emotional lexicon is introduced, to feed the word vectors with emotional seed words. For this step, we used the NRC lexicon [7] to provide the emotional words for the word vectors. The reason for this lexicon's choice is that it provides emotional words in Portuguese and also contains indications for polarities (positive and negative) and annotations for the eight basic emotions according to Plutchik's theory [8], which defines sentiments as *anger, anticipation, disgust, fear, joy, sadness, surprise* and *trust*.

Each word in the emotional lexicon was analyzed in the word vector of each candidate, to identify similarities. For all similar words found with a value higher than 0.7, these words inherited the emotional values from the lexicon's word and - recursively - were analyzed in the word vector to search for new similarities. For each similar word found in the word vectors, we added this word in a "new emotional lexicon" containing the original emotional lexicon and their respective similarities and emotional annotations according to the word vectors.

An important issue to emphasize in this approach is that when finishing the creation of this lexicon, we have a contextual emotional lexicon, because contextual words and its similarities were used in its creation. This context is provided because all messages contain at least one candidate's names. Thus, the context of the lexicon is about politics.

3.2.2 Results

When the lexicon expansion process was finished, the result was a set of personal lexicons about politics, containing the basic lexicon data increased by similarities found in the text and the synonyms of the words. The characteristics of each personal lexicon are presented in Table 1. Due to space limitation, Table 1 only presents the top 5 most known politicians, but in our study, all politicians that were candidate were considered in this analysis.

3.3 Preprocessing

After creating a personal lexicon for each candidate, the next step consisted of creating a text preprocessing pipeline to remove unnecessary information from the texts. This pipeline, as presented in Fig. 3, begins with tokenization, which converts the texts into a list of single words, or *tokens*. Then, the process is divided into two parallel tasks: Part of Speech Tagging (POS-T) and Stopwords Removal. The POS-T process is responsible for identifying grammatical pieces of information for each word in the text, such as adjectives, adverbs and nouns, while the Stopword Removal removes any occurrence in the text of a defined word or list of words.

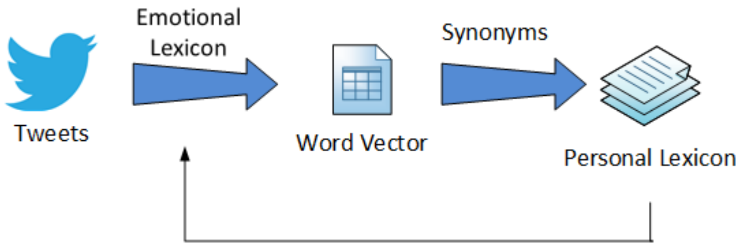

Figure 2: Lexicon expansion process

Table 1: Characteristics of the personal lexicon created

Lexicon	Words	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust	Positive	Negative
Original Lexicon	13911	8,85%	5,92%	7,48%	10,44%	4,88%	8,46%	3,76%	8,72%	16,36%	23,47%
Geraldo Alckmin	15251	8,89%	5,91%	7,50%	10,39%	4,82%	8,45%	3,79%	9,14%	16,56%	23,45%
Jair Bolsonaro	22082	9,16%	5,73%	7,63%	10,50%	5,01%	9,24%	3,38%	10,10%	17,49%	24,11%
Ciro Gomes	15316	8,95%	6,00%	7,47%	10,31%	4,99%	8,65%	3,77%	9,11%	16,78%	23,50%
Fernando Haddad	18264	9,27%	5,70%	7,47%	10,65%	4,87%	8,67%	3,46%	9,19%	17,26%	23,37%
Marina Silva	14842	8,76%	5,88%	7,40%	10,37%	4,85%	8,44%	3,76%	8,79%	16,38%	23,32%

This strategy of paralleling POS-T and Stopwords removal was used because POS-T needs the text in the original format, to classify the words in their respective grammatical categories correctly.

Concerning text cleaning, in POS-T, every word in a grammatical category other than noun, verb, adverb or adjective is discarded. This is important because only these grammatical categories carry emotional information that can be used in further steps. So, more formally, the tokenization process converts the original text D in a set of tokens $T = \{t_1, t_2, ..., t_n\}$ where each element contained in T is part of the original document D. These tokens will feed the POS-T, which will label each token with semantic information. Finally all nouns, verbs, adverbs and adjectives will be collected in a set P, where $P_t = \{p_{(t,1)}, p_{(t,2)}, ..., p_{(t,k)}\}$ and 0 $\leq$ k $\leq$ n and $P_t \subset T$.

The Stopwords list is a manual and predefined set $SW = \{sw_1, sw_2, ...sw_y\}$ of words, intended to avoid the analysis of common and irrelevant words. There are many examples of Stopwords lists on the internet and in libraries for Natural Language Processing (NLP). In our approach, after the Stopwords Removing process, the result list is a set $N = T - SW$.

After the parallel preprocessing tasks finish, the result document ST must contain a set of words where $ST = P \cap N$.

Later, in LM a lemmatizer process reduces the words to their lemma. This step is important because allows considering all inflected words as only one, producing the set of preprocessed texts $PR = \{LM(ST_1), LM(ST_2), ..., LM(ST_z)\}$.

The final result of this pipeline is a new emotional lexicon that considers the similarities of words used in expressions that cite the candidates, and their synonyms, and that is a personal representation of the sentiments about each candidate.

For this preprocessing step, we developed a Python module using Spacy[1] for automatizing the Tokenization, POS-T, Stopwords Removal and Lemmatization processes. We chose to use this toolkit in the development because it provides support for Brazilian Portuguese in all steps described earlier.

3.4 Sentiment Analysis

Once we created new personal lexicons for all candidates, the next step in the dataset creation was to analyze the emotions contained in the texts about each candidate.

For this purpose, we developed a tool that counts the frequency of each emotional word in a text. The result of this analysis is the final dataset, containing the emotional analysis of each Twitter message for each candidate, on a scale from 0 to 100 for each Plutchik's primary emotion.

This approach - a bag-of-word approach - was adopted because we intended to identify which emotions were more relevant to the voters when deciding their candidate, besides to generate a "candidate fingerprint" through the words used to describe them. Moreover, the absence of emotional corpora about politics in Portuguese restricted the possibility of using other techniques to identify the emotions in our texts.

[1] https://spacy.io/

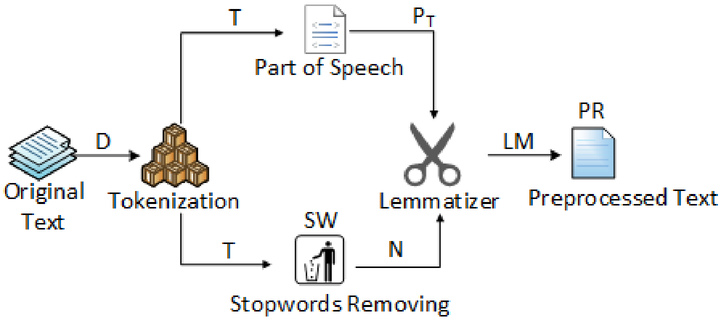

Figure 3: Preprocessing tasks

4 Data analysis

Once the dataset was created, we attempted to use data analysis to identify some particularities about the data, and how these particularities could explain the results of the elections. We used several techniques to identify correlations between the results of the elections and the data analyzed.

Once we identified the emotions that influenced the first-round results and how they did so, the next objective was to predict the results for the second round. To reach this objective, we decided to use an approach based on machine learning. The goal of this approach was to train a model that could accurately predict the percentage of votes for each candidate based on the emotions previously identified, the percentage of votes cast for each candidate in the first round, and the emotions contained in tweets on the day of the second-round vote.

4.1 Training dataset

During the creation of a dataset for training the model, an important issue was identified: how to "translate" the emotions into a percentage of votes. Once the first round results were known, the relationship between candidates' emotional profiles and the percentage of votes cast on that day could also be determined. However, it was necessary to obtain many more examples to train a model. To bypass this obstacle, we chose to use the public information on electoral polls available from public institutes. The chosen institutes were: Instituto Brasileiro de Opinião e Pesquisa (Ibope) [2], Instituto Datafolha [3], Vox Populi [4] and Paraná Pesquisas [5] which are the most important polling institutes in Brazil.

To create the training dataset, we collected the voting intention results of 42 polls for candidates that we had in the dataset, which resulted in 324 examples for training, with 107,039.52 tweets analyzed. Later, knowing the period of each poll, we analyzed the average of each basic emotion for each candidate in the same period from the database. We then transferred these emotions to a new file, indicating the candidate's name, the number of candidates in the poll, period, emotions, and institute.

Despite the number of Tweets messages of each candidate are different, this new dataset contains each candidate's grouped emotions during the period of each poll. Thus, all candidates had the same number of registers in the dataset. This approach ensured that the most cited candidates in messages did not bias the dataset.

4.2 Predicting results

After creating the training dataset, the next step was to train a model for predicting the results for the second round. For this purpose, we analyzed five different machine learning algorithms, to identify the best correlation between data and results. In all cases, the dataset was separated 70% for training and 30% for testing, using the Mean Absolute Error (MAE) as errors standard measure to have a comparison basis between traditional pools and twitter Sentiment Analysis.

The algorithms (using implementations for R and all tuned for the best fit) chosen for the analysis and their results after training the models are presented in Table 2.

Table 2: Algorithms evaluation

Algorithm	Correlation	MAE
Simple Linear Regression	0,2639	10,6302
SVM	0,3677	9,3608
Decision Table	0,5385	9,226
Extreme Gradient Boost	**0,9096**	**0,9787**
Random Forest	0,8088	6,322

The best result was obtained by an Extreme Gradient Boost algorithm, which had a correlation of 0,9096 (very

[2] http://www.ibope.com.br
[3] http://datafolha.folha.uol.com.br/
[4] http://www.voxpopuli.com.br
[5] http://www.paranapesquisas.com.br/

strong correlation) between the results in the polls and the emotions in the same period, and a mean average error of 0,9787.

Once the model was created, the goal was to predict the percentage of votes for each candidate in the second round. In our experiment, we decided to predict the values based only on the emotions expressed on the day of second-round voting until 17:00. This time limitation is because voters could vote only until 17:00. Voters made their decisions based on their emotions during the voting process. Therefore, emotions that were expressed before the second-round election day were not crucial in this analysis.

After collecting the tweets for each second-round candidate - Jair Bolsonaro and Fernando Haddad - on October 20 and analyzing tweets about them using the same process presented in section 3.3, we got the values presented in Table 3.

When applying these values to the model created previously, we got a prediction of **54,58% for Jair Bolsonaro and 43,98% for Fernando Haddad and 0,9787% of MAE** that can be considered as a **correct prevision** because the official results for the second round were 55,15% for Jair Bolsonaro and 44,87% for Fernando Haddad, whose values are in the accepted error margin.

5 Conclusion

Social media have changed the way people interact and express their thoughts about everything. Because of this changing reality and the vast quantity of data available, sentiment analysis is becoming a powerful, fast, and relatively inexpensive tool that is extremely useful for analyzing many different types of scenarios and predicting future results.

Furthermore, although there have been many studies about the influence of social media on elections, there are not approaches using sentiment analysis to identify voters' emotions and predict future election results, while taking into account the results of previous studies.

The correlation between voters' emotions and the percentage of votes shows how vital is to know the audience's sentiments to plan effective strategies for interacting with them. Moreover, the very strong correlations that we found between the basic emotions and poll results', as well our model's successful prediction of the second-round results of Brazilians elections, strongly suggest that sentiment analysis can become a viable and reliable alternative to traditional opinion polls, with the advantages of being much faster and less expensive.

However, it is important to emphasize that there are no studies yet published about what the acceptable threshold is for replacing traditional polls with sentiment analysis. Also, it has not yet been established how many tweets must be analyzed to replace a traditional opinion poll with a sufficient degree of certainty. In future, we plan to identify the relationship between the number of textual opinions and the reliability of our model, to define a threshold for the safe use of sentiment analysis instead of an opinion poll.

Acknowledgements

This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UID/CEC/00319/2019

REFERENCES

[1] Adam Bermingham and Alan Smeaton. On using twitter to monitor political sentiment and predict election results. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, pages 2–10, 2011.

[2] Ronen Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013.

[3] Brian Heredia, Joseph Prusa, and Taghi Khoshgoftaar. Location-based twitter sentiment analysis for predicting the u.s. 2016 presidential election, 2018.

[4] Ricardo Martins, José Almeida, Paulo Novais, and Pedro Henriques. Creating a social media-based personal emotional lexicon. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*, WebMedia '18, pages 261–264, New York, NY, USA, 2018. ACM.

[5] Ricardo Martins, José João Almeida, Pedro Rangel Henriques, and Paulo Novais. Predicting performance problems through emotional analysis (short paper). In *OASIcs-OpenAccess Series in Informatics*, volume 62. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[7] Saif Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[8] Robert Plutchik. Emotions: A general psychoevolutionary theory. *Approaches to emotion*, 1984:197–219, 1984.

Table 3: Emotions in second round's day

Candidate	**Anger**	**Anticipation**	**Disgust**	**Fear**	**Joy**	**Sadness**	**Surprise**	**Trust**
Jair Bolsonaro	13,14%	11,14%	8,18%	12,71%	12,55%	14,80%	6,97%	20,50%
Fernando Haddad	12,39%	13,50%	6,41%	10,67%	14,95%	14,22%	7,57%	20,29%

[9] Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Fourth international AAAI conference on weblogs and social media*, 2010.

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