Analysis of Twitter users' sentiments about the first round 2022 presidential election in Brazil

Daiana Kathrin Santana Santos[1], **Lilian Berton**[1]

[1]Instituto de Ciência e Tecnologia – Universidade Federal de São Paulo (UNIFESP)
12247-014 – São Paulo – SP – Brazil

{daiana.kathrin,lberton}@unifesp.br

Abstract. *The growth of internet and communication through social networks have made it easier to obtain information about what other individuals are thinking and what their opinion is on a given subject, however, a person manually cannot analyze all the comments on the network on a certain topic, requiring the use of technologies, computers and algorithms to assist in data analysis. Therefore, this work aims to collect, process, and classify the feelings of a sample of texts published on Twitter, in Portuguese, about the presidential elections in Brazil in 2022, using the Knowledge Discovery process in Database to analyze the comments and be able to sort the tweets into positive, neutral and negative opinions. We used two classic text representation (Bag of Words and TFIDF) and six classifiers (Naive Bayes, Decision Tree, Random Forest, K-Nearest Neighbors, MLP, and SVM). Thus, predicting which candidate has a greater acceptance/rejection by Brazilians in the 2022 elections, considering only the candidates with the best positions in polls of voting intentions. According to the results obtained using a balanced dataset in the training of algorithms, the candidate with the highest percentage of positive feelings was Jair Bolsonaro, neutral feelings was Luiz Inácio Lula da Silva and negative feelings was Ciro Gomes.*

Resumo. *O crescimento da internet e da comunicação por meio das redes sociais, facilitou a obtenção de informações sobre o que outros indivíduos estão pensando e qual a opinião deles para determinado assunto, porém, uma pessoa manualmente não consegue analisar todos os comentários na rede sobre certo tema, sendo necessário o uso de tecnologias, computadores e algoritmos para auxiliar na análise dos dados. Diante disso, este trabalho tem o objetivo de coletar, processar e classificar os sentimentos de uma amostra de textos publicados no Twitter, em Português, sobre as eleições presidenciais no Brasil em 2022, utilizando o processo de Descoberta de Conhecimento em Base de Dados para analisar os comentários e conseguir classificar os tweets entre opiniões positivas, neutras e negativas. Nós usamos duas representações textuais clássicas (Bag of Words and TFIDF) e seis classificadores (Naive Bayes, Decision Tree, Random Forest, K-Nearest Neighbors, MLP, e SVM) Desse modo, predizer qual candidato tem uma maior aceitação/rejeição por parte dos brasileiros nas eleições de 2022, considerando apenas os candidatos com melhores posições nas pesquisas de intenção de votos. De acordo com os resultados obtidos empregando um dataset balanceado no treinamento dos algoritmos, o candidato com maior porcentagem de sentimentos positivos foi Jair Bolsonaro, sentimentos neutros foi Luiz Inácio Lula da Silva e sentimentos negativos foi Ciro Gomes.*

1. Introduction

The use of the internet and social media to share opinions is becoming more and more common. According to the CGI research [1], in 2020 there were 152 million Brazilians aged 10 or over using the Internet, corresponding to 81% of the Brazilian population; a considerable number of people, enough to determine the result of an election, for example.

Society makes several day-to-day decisions based on the opinions of close people and, with the growth in the use of social networks to share information, many individuals are influenced in their opinions [de Camargo Penteado and Guerbali 2016]. The demonstrations against President Dilma Rousseff, in 2015, on Twitter, had great importance in the contemporary political agenda, in addition to the ability to mobilize and publicize protest acts.

Twitter is a social network, launched in 2006, which promotes a public and secure conversation between its users [Jack Dorsey]. In 2020, there were around 14.1 million Twitter users in Brazil, registering around 500 million "*tweets*" per day [2] that generate great repercussions inside and outside the network. Several artists, companies, prominent people, and politicians have Twitter accounts, with thousands of followers. For example, in 2023 the current president of Brazil, Lula, has 7.9 million, while ex-president Bolsonaro has 11.4 million followers.

Sentiment analysis is a recent area of computer science, which studies the classification of emotions and opinions in texts [Koehn and Mihalcea 2009, Garcia and Berton 2021]. Many users who seek information, opinions, and experiences about a product, before making a purchase, perform sentiment analysis in an intuitive way. A technique widely used to perform sentiment analysis is data mining, which consists of exploring a database using appropriate algorithms, in order to obtain knowledge [Ferrari and Silva 2017]. Data mining is one of the steps of KDD (*knowledge discovery in databases*), consisting of 3 general operations: Pre-processing (selection, organization, and treatment of the database), Mining, and Post-processing (facilitating the interpretation and evaluation of results obtained in mining) [Passos and Goldschmidt 2005].

Data mining corresponds to the application of algorithms to obtain knowledge from data. It is an interdisciplinary and multidisciplinary area that involves knowledge of areas such as databases, statistics, machine learning, and among others [Ferrari and Silva 2017]. Machine Learning uses powerful tools for discovering knowledge in databases, it is an area of artificial intelligence that teaches computers to learn from past experiences [Lorena et al. 2000]. One of the machine learning techniques is classification, which consists of discovering a way to label a set of data from predefined definitions.

Considering all these concepts, this work aims to analyze the feelings of a sample of texts published on Twitter, using the KDD process and machine learning algorithms to analyze the data and classify the posts between positive and neutral, and negative opinions, thus predicting which candidate, among the best placed in polls of voting intentions, has greater acceptance/rejection by Brazilians in the 2022 elections. We employed two classic text representation (Bag of Words and TFIDF) and six classifiers (Decision Tree,

[1] https://www.cgi.br/noticia/releases/cresce-o-uso-de-internet- during-the-pandemic

[2] `https://www.oberlo.com.br/blog/estatisticas-twitter`

Naive Bayes, Random Forest, SVM, MLP and K-NN). The results showed that SVM and Random Forest achieved the highest results with TFIDF vectorization.

The main contributions of the work are: 1) Determine the positive, neutral, and negative sentiments of tweets in Portuguese about the presidential election in Brazil in 2022; 2) Compare the performance of different classifiers in classifying tweets; 3) Analyze whether the textual vectorization can result in different results by the classifiers; 4) Compare computational analyzes with official election results.

The remaining of the paper is organized as follows. Section 2 presents the related work. Section 3 presents the methodology where these steps were executed: the obtaining and analyzing data; preprocessing; manual classification of tweets and automatic classification. Section 4 presents the automatic classification results. Section 5 presents conclusion and future works.

2. Related work

Previous works analyzed feelings of *tweets* related to Brazilian politics, below some of them are mentioned. These works employed only one algorithm to classify the results while we compared six classifiers. As far as we know no previous work analyzed the 2022 elections in Brazil.

In [Silva 2018], a sentiment analysis applied to politics was carried out, in which versions of the SVM and logistic regression were implemented to explore the network of Twitter users to analyze the feelings of its users about the 2018 political scenario in Brazil. The author obtained an average accuracy of 71% and a suggestion for improvement would be the possibility of more users evaluating the same set of texts to guarantee a consensus when associating an opinion with a text.

In [de Queiroz and Almeida 2020], sentiment analysis is applied to *tweets* posted by the five best-ranked candidates in the polls of voting intentions in the first round of the 2018 Brazilian presidential elections. The *tweets* were obtained using the API and subjected to pre-processing techniques, the use of lexical dictionaries, TFIDF vectorization, and the *algorithm K-means* for grouping data into 2 classes, positive and negative. As a recommendation for future work, the use of new algorithms was mentioned, since *K-means* is not always the best option and there are more specific algorithms for sentiment analysis. The second recommendation was to apply the new algorithms to discover new feelings and not just the polarity of the content.

[Viana 2014] tries to gauge the feeling of users from Twitter in relation to the election of 3 candidates for the presidency of Brazil in 2014. Classifying the messages as positive, negative, neutral, and ambiguous with the Naive algorithm Bayes. The candidates Dilma Rousseff, Marina Silva, and Aécio Neves were considered, with an analysis of the feelings for each candidate. One suggestion by the author for future work would be to use another classifier, in addition to Naive Bayes.

In [Caetano et al. 2017], an analysis of political homophily among Twitter users during the 2016 American presidential election was proposed. About 3.6 million *tweets* were collected over 122 days in relation to the candidates, Donald Trump and Hillary Clinton. The *SentiStrength* tool was used, which returns three values between positive, negative, and *scale* (difference between positive and negative values). The results showed

that there is greater homophily among users who share negative feelings. As a suggestion for future work, is the addition of new features such as *hashtags* for classifying users' political discourse on Twitter.

3. Methodology

The execution of this project was divided into 5 phases: obtaining and analyzing data; preprocessing; manual classification of *tweets* and finally automatic classification. The codes were created in the *Jupyter Nootebok* software, with the *Python* language and run on an HP notebook, intel core I5 and 8GB of RAM memory. The source code is stored on the GitHub platform [3].

3.1. Getting data

The data was obtained using the Twitter API, which allows the developer to have access to all posts published on the network, in an easier way. So, the first step for the collection was to make the connection with the API and, to have access to the posts, it was necessary to determine specific keywords or accounts.

Since there is a limit of 2 million *tweets* per month per user connected to the API, it was chosen to analyze only the 4 candidates who were most popular in the XP/Ipespe survey [Alencar], carried out on the day July 25, 2022, which named Lula, Jair Bolsonaro, Ciro Gomes and Simone Tebet as the most popular. Due to this collection limit, it was also chosen to analyze specific periods of the elections. Since it would not be possible to analyze the entire electoral period, the first days that the official candidates were released (00:00 on August 16 to 00:00 on August 22, 2022) and the last days before the 1st round of elections (18:00 from the 28th of September until 5:00 pm on the 2nd of October 2022).

To filter the *tweets* that would be of interest to the project, keywords were used, such as the names of political parties (PT, PL, PDT and MDB); the names/nicknames of the presidential candidates (Lula, Luiz Inácio, Alckmin, Bolsonaro, Braga Netto, Ciro, Ana Paula Matos, Simone Tebet, Mara Gabrilli) and among other words associated with the theme (election and voting).

The API also imposes a collection limit of 100 *tweets* per request. Therefore, it was decided to do one per candidate every minute, with the aim of being able to collect all the existing texts on the network. Algorithm 1 demonstrates how the initial electoral period was collected.

3.2. Data analysis

After collecting the texts about the 4 candidates, some analyzes were made to draw previous conclusions about the data, to find out the total and hourly amount of *tweets* per candidate and what were the words or *hashtags* most common in the collected texts.

In the end, about 2 million *tweets* were collected, with the amount of each candidate shown in Table 1.

After obtaining the data, it was analyzed how many *tweets* per hour each candidate was mentioned, the result is shown in Figure 1.

[3] https://github.com/DaianaKathrin/University-UNIFESP/tree/master/TCC-AnaliseDeSentimentos

Algorithm 1 Getting Data

$TokensLula \leftarrow (lula)\ or\ (PT\ eleic)\ or\ (luiz\ inacio)\ or\ (alckmin)$
$TokensBolsonaro \leftarrow (bolsonaro)\ or\ (PL\ eleic)\ or\ (braga\ netto)$
$TokensCiro \leftarrow (ciro)\ or\ (PDT\ eleic)\ or\ (Ana\ Paula\ Matos)$
$TokensSimone \leftarrow (simone\ tebet)\ or\ (MDB\ eleic)\ or\ (Mara\ Gabrilli)$

while $numberTweets \leq 2000000$ **do**
$EndTime + 1minute$

$tweets \leftarrow BuscaTweets(LulaTokens, StartTime, EndTime)$
$tweetsLula \leftarrow$ *tweets*$Lula +$ *tweets*

$tweets \leftarrow BuscaTweets(BolsonaroTokens, StartTime, EndTime)$
$tweetsBolsonaro \leftarrow$ *tweets*$Bolsonaro +$ *tweets*

$tweets \leftarrow BuscaTweets(CiroTokens, StartTime, EndTime)$
$tweetsCiro \leftarrow$ *tweets*$Ciro +$ *tweets*

$tweets \leftarrow BuscaTweets(TokensSimone, StartTime, EndTime)$
$tweetsSimone \leftarrow$ *tweets*$Simone +$ *tweets*

$StartTime \leftarrow EndTime$
end while

Table 1. Number of *tweets* collected per candidate

	Lula	Bolsonaro	Simone	Ciro
Number of *tweets* collected	931,730	946,438	131,162	498,066

The number of mentions of candidates Bolsonaro and Lula were similar throughout the period. In the period 16 August to 29 September, there was much less data than in the period 30 September to 2 October, probably caused by the closer approach to polling day (2 October) and the debate in the television station Globo, made on the 29th of September. We noticed that the increase in *tweets* started at the end of the 29th in the same period that the debate took place.

Figure 2 shows which were the most common hashtags among all candidates and the most used was "#DebateNaGlobo", which indicates that the debate influenced the number of *tweets*.

In the official result of the 1st round of the elections, Lula got 48.43% of the votes, Bolsonaro got 43.20%, Simone Tebet got 4.16% and Ciro got 3.04%. Figure 3 shows the comparison between the percentages of the official vote for the 2022 presidential election in Brazil and the number of *tweets* collected for each candidate.

Figure 1. Number of *tweets* per hour and per candidate

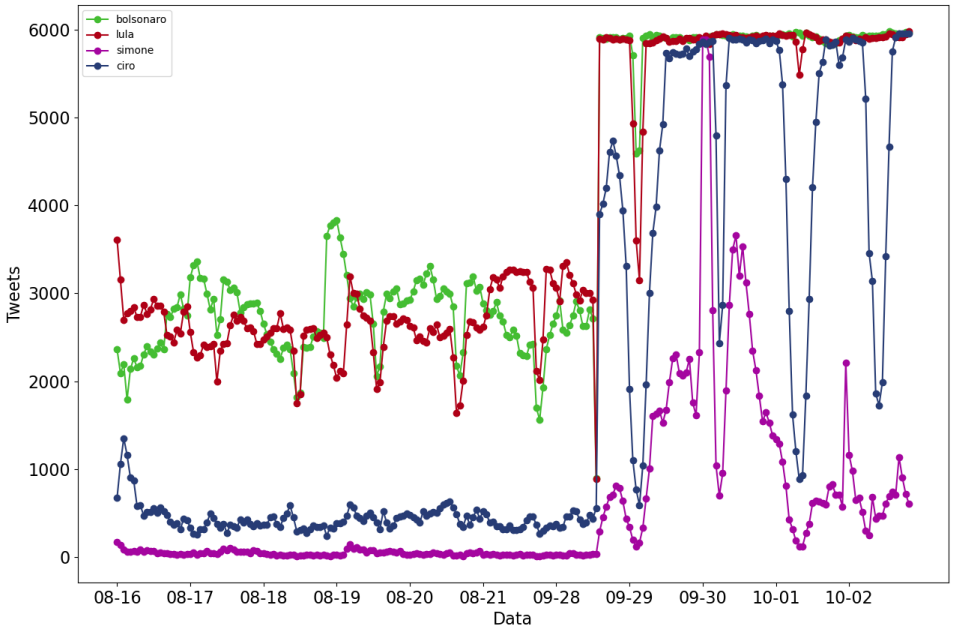


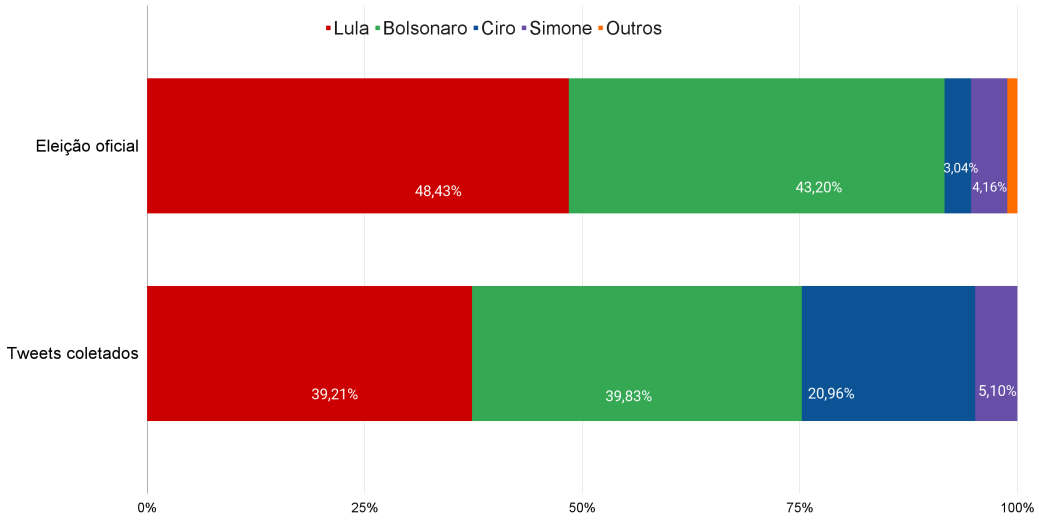
Figure 2. Most used *hashtags* considering all candidates

3.3. Pre-Processing

In this phase, the data is cleaned, leaving only what is relevant to the study. Therefore, *stopwords*, meaningless words such as articles, prepositions, and conjunctions that hardly characterize documents, were removed [Matsubara et al. 2003]. In addition, stemming was also performed and punctuations and accents that did not add value to the results were removed.

After processing the words, the text was separated into *tokens* and applied the

Figure 3. Comparison between official elections and collected *tweets*



techniques of Bag of Words (BW) which measure the occurrence of words in the text, and Term-Frequency Inverse Document Frequency (TFIDF) which measures the frequency with which a term occurs in a document to represent the data.

3.4. Manual classification of *tweets*

The texts obtained in the posts are classified into three classes, positive, neutral, and negative. Each candidate has its subset and the feelings are analyzed considering the candidate in question, for example, if there is a tweet speaking well of candidate X and badly of candidate Y, in the dataset of person X, the text will be set to positive and on the dataset of candidate Y as negative. Table 2, shows an example of a comment for each candidate and associated sentiment.

To manually classify the sentiments of some tweets, we randomly separated 20% of the base and classified some tweets during a period. Because it is massive and manual work, it demands a lot of time and attention. The author of the work and 3 different people (family members of the author, with no experience in academic research), guided by the author, performed the manual classification. The texts can be considered subjective because sometimes they are ironic speeches and/or have a context in the past, therefore, it will depend on the interpretation of who is reading.

To increase the number of classified *tweets* and as suggested by authors in past works, some *hashtags* representative of the feelings for each candidate were used to automatically classify more *tweets* in the rest of the set. This approach was applied after having manually classified as much as possible.

3.5. Automatic classification

Using the results of the manual classifications, the classification algorithms are trained to understand the behavior and find patterns in the collected *tweets*. Since no algorithm is the best and with 100% confidence to classify all texts correctly, 6 different algorithms

Table 2. Sample *tweets* for each candidate and corresponding sentiment

	Positive	**Neutral**	**Negative**
Bolsonaro	”@TerraBrasilnot É agora que queremos vê se a lei vale para todos. Afinal existe a lei da ficha limpa. Jair Messias Bolsonaro o melhor Presidente que nossa nação já teve e terá até 31/12/26 O capitão do povo #BolsonaroReeleitoEm2022 #CapitaoDoPovoQue VaiVencerDenovo”	”Nova pesquisa nacional, com amostragem mais ampla, põe Bolsonaro a apenas 4 pontos da vitória no primeiro turno”	”Tchutchuca do centrão! Bolsonaro ladrão! #ForaBolsonaro #tchutchucaDoCentrao”
Lula	”Até ele faz o L! É LULA primeiro turno! FAZ O L! FAZ O L! FAZ O L! FAZ O L! FAZ O L! FAZ O L! FAZ O L! FAZ O L! FAZ O L! FAZ O L! FAZ O L! FAZ O L! #LulaNo1ºTurno #LulaNo1oturno #LulaNoDebate #LulaNoPrimeiroTurno #LulaNoPrimeiroTurno13 ”	”O que Ciro, Bolsonaro, Lula e Tebet propõem para a economia”	”Uma afirmação difícil de refutar: Lula é ladrão.”

	Positive	**Neutral**	**Negative**
Simone Tebet	”Pra vc trabalhador que é fazendeiro e do agro negócio a Simone Tebet é muito preparada.”	”PRO DIA NASCER FELIZ! @LulaOficial LULA JÁ VENCEU NO PRIMEIRO TURNO NA NOVA ZELÂNDIA Lula %: 72.9% Bolso%: 15.74% Lula 329 Bolso 71 Ciro 23 Padre 3 Simone tebet 8 Felipe 14 Leo 2 Sofia 1 Total 451”	”Que bom que Simone Tebet não vai ganhar Porque ia ser bem feio pra ela quando ela falasse que não ia ser possível dar os 5 mil reais pra todo jovem que concluísse o ensino médio”
Ciro	”#VoteNoTerceiro #CiroNoSegundoTurno SÓ CIRO VENCE LULA É CIRO CONTRA O SISTEMA! EU ESTOU FECHADO COM CIRO”	”Resultado da Votação aqui na Dinamarca: - Lula 936 - Bolsonaro 147 - Ciro 58 #Eleicoes2022”	”Alguns ainda gostam do Ciro porque ele parece estar vivo, mas ele é vivo e mentiroso: de cada 10 coisas que o Ciro diz, 8 são mentiras.”

were used to classify the data (Naive Bayes, Decision Tree, Random Forest, K-Nearest Neighbors, MLP, and SVM).

The algorithms were used with the *scikit-learn* library of the Python language, in version 1.0.2 and the chosen parameters are presented as follows:

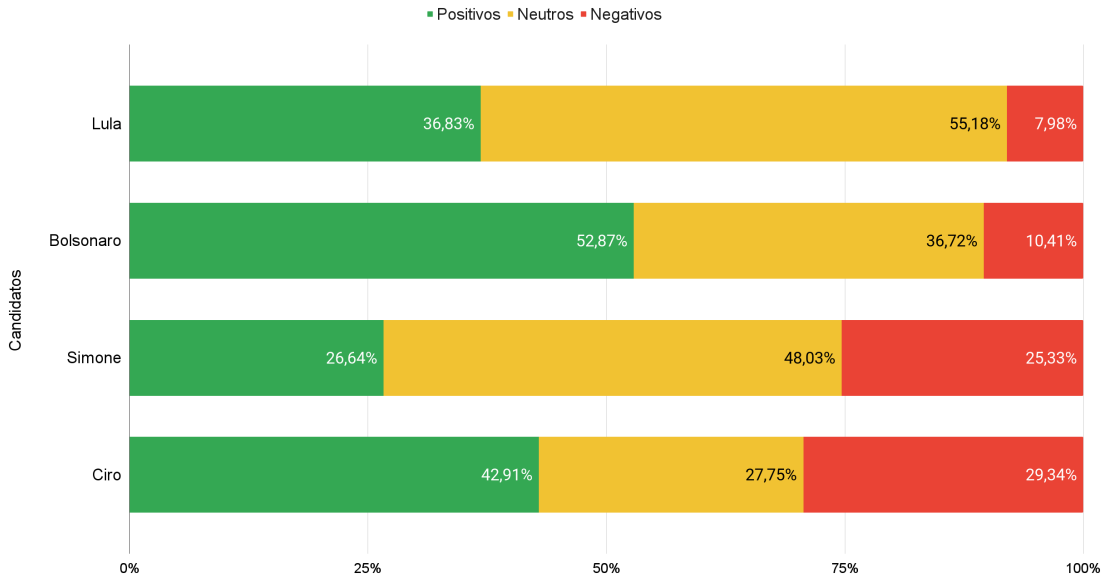
- Naive Bayes:
  - *priors*= $None$
  - *var_smoothing*= $1e-09$
- Decision Tree:
  - *criterion*=$'gini'$
  - *splitter*=$'best'$
  - *max_depth*= $None$
  - *min_samples_split*= 2
  - *min_samples_leaf*= 1
  - *min_weight_fraction_leaf*= 0.0
  - *max_features*= $None$
  - *random_state*= $None$
  - *max_leaf_nodes*= $None$
  - *min_impurity_decrease*= 0.0
  - *class_weight*= $None$
  - *ccp_alpha*= 0.0
- *Random Forest*:
  - *n_estimators*= 100
  - *criterion*=$'gini'$
  - *max_depth*= $None$
  - *min_samples_split*= 2
  - *min_samples_leaf*= 1
  - *min_weight_fraction_leaf*= 0.0
  - *max_features*=$'sqrt'$
  - *max_leaf_nodes*= $None$
  - *min_impurity_decrease*= 0.0
  - *bootstrap*= $True$
  - *oob_score*= $False$
  - *n_jobs*= $None$
  - *random_state*= $None$
  - *verbose*= 0
  - *warm_start*= $False$
  - *class_weight*= $None$
  - *ccp_alpha*= 0.0
  - *max_samples*= $None$
- K-NN:
  - *n_neighbors*= 5
  - *weights*=$'uniform'$
  - *algorithm*=$'auto'$
  - *leaf_size*= 30
  - *p*= 2
  - *metric*=$'minkowski'$
  - *metric_params*= $None$
  - *n_jobs*= $None$
- SVM:
  - *C*= 1.0
  - *kernel*=$'rbf'$
  - *degree*= 3
  - *gamma*=$'scale'$
  - *coef0*= 0.0
  - *shrinking*= $True$
  - *probability*= $False$
  - *tol*= 0.001
  - *cache_size*= 200
  - *class_weight*= $None$
  - *verbose*= $False$
  - *max_iter*= -1
  - *decision_function_shape*=$'ovr'$
  - *break_ties*= $False$
  - *random_state*= $None$
- MLP:
  - *hidden_layer_sizes*= 100
  - *activation*=$'relu'$
  - *solver*=$'adam'$
  - *alpha*= 0.0001
  - *batch_size*=$'auto'$
  - *learning_rate*=$'constant'$
  - *learning_rate_init*= 0.001
  - *power_t*= 0.5
  - *max_iter*= 400
  - *shuffle*= $True$
  - *random_state*= $None$
  - *tol*= 0.0001
  - *verbose*= $False$
  - *warm_start*= $False$
  - *momentum*= 0.9
  - *nesterovs_momentum*= $True$
  - *early_stopping*= $False$
  - *validation_fraction*= 0.1
  - *beta_1*= 0.9
  - *beta_2*= 0.999
  - *epsilon*= $1e-08$
  - *n_iter_no_change*= 10
  - *max_fun*= 15000

We performed the experiment with balanced data (95 tweets per class). All the previously mentioned classification algorithms were trained and a prediction function called *predict_proba()* was used, which provides the probability of each class being chosen, hence for conclusion the sentiment is chosen with greater probability, summing the result of all algorithms.

4. Results

After classifying the *tweets*, Bolsonaro achieved the highest number of positive comments with 500,381, followed by Lula 343,190, Ciro 213,701, and Simone 32,280. In the negative texts, Ciro leads with 146,145, then Bolsonaro 98,558, Lula 74,396, and Simone 40,686. While in the neutral comments in first place was Lula 514,145, after Bolsonaro 347,499, Ciro 138,220, and Simone 58,196. The percentage comparison of each sentiment for each candidate is shown in Figure 4.

Figure 4. Percentages of predicted positive, neutral, and negative feelings for each candidate

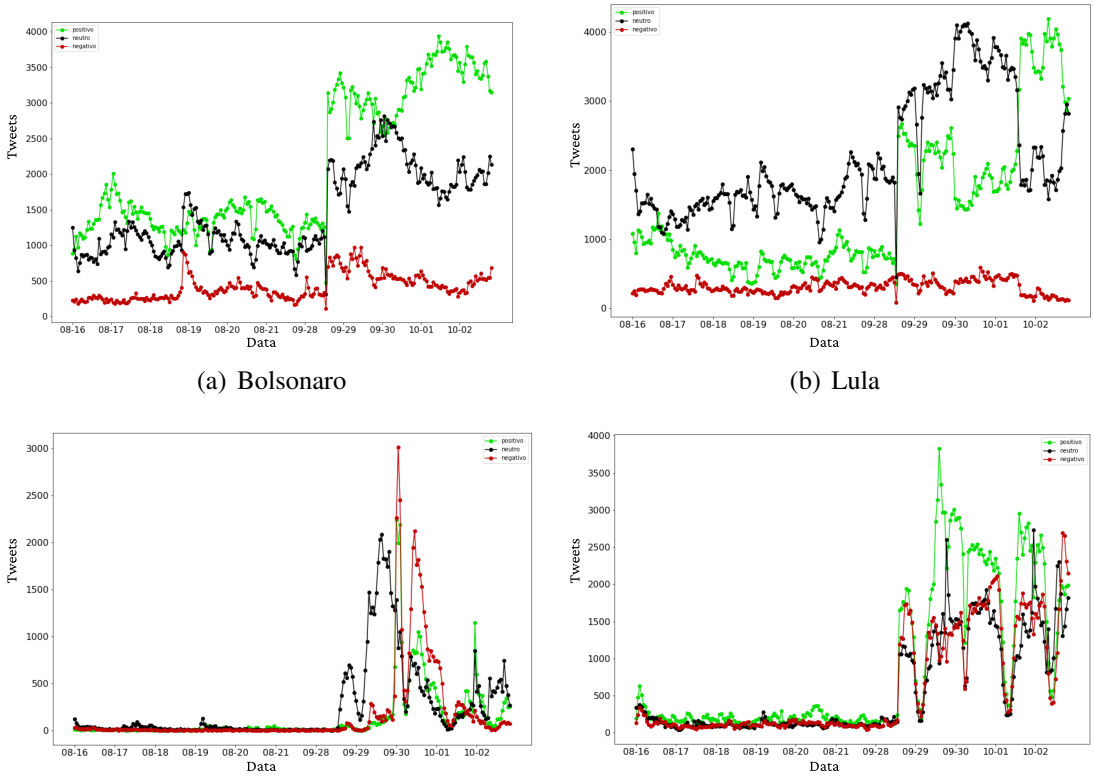


In order to analyze whether there was a time when the number of feelings about a candidate stood out, a graphic was constructed for each candidate in Figures 6(a), 6(b),6(c) and 6(d) where there are 3 lines about the respective positive, neutral and negative feelings.

Analyzing the images, all had an increase in the number of comments when there was a debate between the candidates, broadcast on national television. In the figure 6(a) for candidate Bolsonaro, most of the time there were more positive comments, except for a moment on August 19, when the number of neutral texts exceeded the positive ones and it was also the section where there was a peak of negative comments. Over the hours, candidate Bolsonaro was the one with the most negative and positive comments, when compared to the others.

When we analyze Figure 6(b) that presents the comments of candidate Lula, it has the highest number of neutral comments over time. On October 2, there was an increase in

Figure 5. Positive, negative, and neutral hourly comments predicted for each candidate



(a) Bolsonaro

(b) Lula

(c) Simone

(d) Ciro

positive comments and a decrease in neutral and negative comments, which was a positive point for him, since it was the voting day.

Candidates Ciro and Simone had a smaller number of comments over time, compared to Bolsonaro and Lula. At the end of the 29th of September, the number of positive comments from Ciro increased, while in the same period, the number of negative comments from Simone also increased, perhaps this may have had something conneted.

Table 3 shows the evaluation of each algorithm using the metrics Accuracy, Precision, Recall and f1-score in the tests performed by the Decision Tree, Naive Bayes, Random Forest, SVM, MLP and K-NN using BW and TFIDF vectorization techniques. The results generated by the metrics varied between 15% and 70% with SVM and Random Forest achieving the highest results. All algorithms that presented the highest percentages were in the TFIDF vectorization, which does not mean that it is better or worse than BW, but that it performed better in this problem.

Table 3. Classification results

Bolsonaro		Accuracy	Precision	Recall	f1-score
BW	Decision Tree	37,69%	39,56%	58,54%	28,23%
	Naive Bayes	37,52%	37,85%	62,48%	26,50%
	Random Forest	56,39%	39,65%	67,12%	34,92%
	SVM	60,37%	38,21%	52,75%	35,08%
	MLP	47,46%	36,91%	48,18%	30,02%
	K-NN	15,85%	35,61%	56,47%	12,99%
TFIDF	Decision Tree	53,48%	39,39%	48,08%	33,74%
	Naive Bayes	46,25%	36,80%	63,16%	29,54%
	Random Forest	63,84%	40,61%	50,79%	36,86%
	SVM	64,68%	38,78%	71,14%	36,78%
	MLP	48,92%	37,66%	64,83%	31,05%
	K-NN	39,90%	36,84%	62,98%	27,06%
	Mean	47,70%	38,16%	58,88%	30,23%

Lula		Accuracy	Precision	Recall	f1-score
BW	Decision Tree	42,45%	38,25%	61,15%	29,27%
	Naive Bayes	35,43%	36,19%	57,37%	25,38%
	Random Forest	47,43%	41,00%	43,20%	31,78%
	SVM	46,15%	39,79%	59,66%	30,90%
	MLP	50,30%	38,43%	47,53%	32,08%
	K-NN	29,78%	36,91%	51,97%	22,28%
TFIDF	Decision Tree	49,43%	39,76%	62,58%	32,48%
	Naive Bayes	50,86%	36,33%	62,31%	30,73%
	Random Forest	51,35%	40,99%	62,83%	33,56%
	SVM	54,84%	41,17%	63,17%	34,23%
	MLP	53,11%	38,67%	64,35%	33,05%
	K-NN	50,36%	38,63%	66,00%	32,50%
	Mean	46,79%	38,84%	58,51%	30,69%

Simone		Accuracy	Precision	Recall	f1-score
BW	Decision Tree	40,00%	51,92%	60,41%	35,62%
	Naive Bayes	61,48%	50,55%	58,55%	46,87%
	Random Forest	65,92%	51,98%	61,58%	52,15%
	SVM	57,77%	54,81%	71,11%	47,09%
	MLP	59,25%	51,65%	56,32%	46,51%
	K-NN	68,14%	49,43%	45,01%	45,38%
TFIDF	Decision Tree	65,92%	52,74%	76,70%	54,66%
	Naive Bayes	60,74%	52,70%	74,21%	48,59%
	Random Forest	70,37%	55,84%	79,64%	57,24%
	SVM	61,48%	56,74%	73,25%	49,03%
	MLP	61,48%	51,07%	74,25%	49,46%
	K-NN	67,40%	56,25%	78,12%	53,77%
	Mean	61,66%	52,97%	67,43%	48,86%

Ciro		Accuracy	Precision	Recall	f1-score
BW	**Decision Tree**	26,99%	33,51%	48,68%	14,56%
	Naive Bayes	39,32%	33,59%	61,81%	19,40%
	Random Forest	38,45%	33,47%	62,79%	18,95%
	SVM	24,70%	33,74%	40,25%	14,07%
	MLP	43,22%	33,51%	65,66%	20,66%
	K-NN	19,78%	33,32%	39,88%	11,26%
TFIDF	**Decision Tree**	50,57%	33,57%	66,85%	22,98%
	Naive Bayes	52,67%	33,64%	65,00%	23,70%
	Random Forest	51,34%	33,49%	60,72%	23,11%
	SVM	52,07%	33,80%	67,35%	23,80%
	MLP	52,20%	33,61%	67,39%	23,58%
	K-NN	37,16%	33,43%	58,53%	18,44%
	Mean	40,71%	33,56%	58,74%	19,54%

5. Conclusion

The objective of analyzing the opinions of Brazilians about the 2022 elections on the social network Twitter using machine learning techniques was accomplished. The official result of the 1st round of elections, Lula got 48.43%, Bolsonaro got 43.20%, Simone Tebet got 4.16% and Ciro got 3.04% of the votes. According to our analyses, the total percentage of *tweets* for each candidate was: Bolsonaro with 37.75%, Lula with 37.16%, Ciro with 19.86%, and Simone with 5.23%. The candidate who had the most positive comments was Bolsonaro, the most neutral was Lula and the most negative was Ciro. No relationship was observed with the official result that could indicate a prediction, but the polarization between the candidates, Lula and Bolsonaro, being the most commented and were the most voted, with similar amounts.

Some improvements can be made in future work. The ideal would be to manually classify a larger number of texts, thus increasing confidence in the results. Another improvement would be to be able to analyze the entire election period, from the official launch of the candidates to the voting time, however, as it generates an excess of data, it is necessary to separate a large and reliable space for storage and get an account with more privileges on Twitter API. The third improvement would be to analyze other social networks besides Twitter. Changing algorithm parameters can bring relevant results, as well as trying new methods, and using other text vectorization models like *word embeddings*.

References

[Alencar] Alencar, C. Xp/ipespe: após intervalo de 52 dias, lula tem 44%, e bolsonaro, 35%. [On-line], `https://www.uol.com.br/eleicoes/2022/07/25/pesquisa-xp-ipespe-presidente-julho.htm`. accessed: 26.07.2022.

[Caetano et al. 2017] Caetano, J. A. C., Lima, H. S. L., dos Santos Santos, M. F., and Marques-Neto, H. T. M.-N. (2017). Utilizando análise de sentimentos para definição da homofilia política dos usuários do twitter durante a eleição presidencial americana de 2016. In *Congresso da Sociedade Brasileira de Computação-CSBC*.

[de Camargo Penteado and Guerbali 2016] de Camargo Penteado, C. L. and Guerbali, J. G. (2016). As manifestações do impeachment no twitter: uma análise sobre as manifestações de 2015. *Ponto-e-Vírgula: Revista de Ciências Sociais*, (19).

[de Queiroz and Almeida 2020] de Queiroz, G. G. and Almeida, L. (2020). Uma metodologia de análise de sentimentos dos candidatos as eleições presidenciais de 2018 no twitter. *Revista de Engenharia e Pesquisa Aplicada*, 5(1):21–30.

[Ferrari and Silva 2017] Ferrari, D. G. and Silva, L. N. d. C. (2017). *Introdução a mineração de dados*. Saraiva Educação SA.

[Garcia and Berton 2021] Garcia, K. and Berton, L. (2021). Topic detection and sentiment analysis in twitter content related to covid-19 from brazil and the usa. *Applied soft computing*, 101:107057.

[Jack Dorsey] Jack Dorsey, E. W. e. B. S. Sobre o twitter. [On-line], `https://about.twitter.com/`. accessed: 27.01.2022.

[Koehn and Mihalcea 2009] Koehn, P. and Mihalcea, R. (2009). Proceedings of the 2009 conference on empirical methods in natural language processing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*.

[Lorena et al. 2000] Lorena, A. C., Gama, J., and Faceli, K. (2000). *Inteligência artificial: uma abordagem de aprendizado de máquina*. Grupo Gen-LTC.

[Matsubara et al. 2003] Matsubara, E. T., Martins, C. A., and Monard, M. C. (2003). Pretext: Uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words. *Techinical Report*, 209(4):10–11.

[Passos and Goldschmidt 2005] Passos, E. and Goldschmidt, R. (2005). Data mining: um guia prático. *Editora Campus, Rio de Janeiro*.

[Silva 2018] Silva, L. R. (2018). Análise de sentimentos aplicada à política.

[Viana 2014] Viana, Z. L. (2014). Mineração de textos: análise de sentimentos utilizando tweets referentes às eleições presidenciais 2014.