Artigo

em questão
E-ISSN 1808-5245

# Who will be evicted from Big Brother Brazil? Forecasting based on social networking

**William Takahiro Maruyama** [I]
https://orcid.org/0000-0001-6445-0666

**Luciano Antonio Digiampietri** [I]
https://orcid.org/0000-0003-4890-1548

[I] Universidade de São Paulo, São Paulo, SP, Brasil

**Abstract:** The advent of social media has provided individuals with a platform for the expression of opinions and preferences on an extensive range of subjects. The data obtained from social media can be a valuable source of information for the analysis of audience intentions and interests, thereby facilitating the process of informed decision-making and strategy development. One example of human behavior analysis is the examination of popular voting events. In order to address the challenge of forecasting a sequence of events in the context of voting, we propose a novel methodology that employs a data-driven solution, incorporating Twitter/X data and regression models. In this case study, we employed techniques commonly utilized in electoral outcome forecasting, including volumetric and sentiment analysis, to predict the eviction of contestants in Big Brother Brazil. Our experiments resulted in an average absolute error of approximately 11, with an accuracy of 81.25% for predicting evictions and 68.75% for forecasting the classification order.

**Keywords**: forecasting; Big Brother Brasil; sentiment analysis; volumetric analysis; social network analysis

## 1 Introduction

The Brazilian adaptation of Big Brother, commonly referred to as Big Brother Brasil (BBB), presents a unique context for examining social dynamics and decision-making under controlled conditions. In this program, a group of individuals, known as Houseguests, are confined to a house under 24/7 surveillance, without external influences, and tasked with the objective of remaining in the house until the final day to compete for the final prize. The crucial step of the BBB program is the Nomination. Typically, the fellow Houseguests vote to decide who will participate in eviction from the house. In the eviction phase, which is referred to as "Paredão" in Portuguese, the decision is made by the audience through a majority vote.

Social media has become an increasingly influential platform for individuals and organizations to share their thoughts and opinions, influence others, and connect. The social media platform X, previously known as Twitter, offers a dynamic digital environment where many people can discuss various subjects. Some public figures, such as the BBB contestants, use these platforms by leveraging them to communicate, influence, and connect with large audiences.

The advent of social media has led to the creation of a vast expanse of publicly available data, which presents a unique opportunity to explore the intricacies of human behavior. Social media platforms where users naturally express their opinions provide a rich source of data that promises many applications. Researchers can use these data for audience behavior analysis, trend identification, and predictive modeling (Rousidis; Koukaras; Tjortjis, 2019). By making use of publicly accessible data, researchers can develop and evaluate computational models to anticipate the results of voting processes.

In this article, we propose a new methodology for predicting the outcomes of popular events. Given the potential of social networks and the popularity of BBB, this experiment aims to forecast which Houseguests will be evicted from the house based on texts published by users.

We consider each eviction as a discrete voting process with a brief time frame, and our proposed methodology represents a preliminary investigation into the potential for predicting new election results based on previous outcomes. We obtained a substantial data set from Twitter/X, consisting of over one million tweets. Related work attempts to predict the outcome of a single election (Brito; Silva Filho; Adeodato, 2021). Several studies have focused on one or more of the following approaches: sentiment analysis, volumetric analysis, or machine learning. In this study, we integrated these approaches and employed the stance identified through hashtag analysis. Furthermore, we present a strategy for output normalization and two metrics for evaluating the performance of the models, which are not limited to Mean Absolute Error (MAE).

Professionals engaged in public engagement and celebrity career management can benefit from this more efficient and effective approach. The forecasts can assist participants in formulating more effective and targeted actions for their campaigns. The findings of this research have the potential to contribute to different academic disciplines, including marketing, political science, psychology, and sociology. Our findings illustrate the value of computational models in gaining a deeper understanding of public opinion.

## 2 Related works

Online platforms serve as valuable resources for researchers analyzing and forecasting various applications. Academic literature offers readily accessible applied research utilizing different social media platforms in various countries (Soler; Cuartero; Roblizo, 2012). This study aims to forecast evictions in BBB using concepts and techniques inspired by research on predicting election outcomes.

In the same context, Figueredo, Marinho and Alves (2015) conducted a study to investigate the correlation between participants' performance and their Twitter popularity. The study analyzed two specific television shows, Superstar 2014 and Big Brother Brasil 2015. The authors created a proprietary data collection tool for obtaining tweets. They utilized both Twitter's search engine and emoticons to determine the polarity of sentiment. The authors collected a total of 193,415 tweets for BBB and 9,938 tweets for Superstar. After conducting a temporal analysis of Superstar, the authors observed that the total number of tweets aligns with the program's result. The researchers identified a correlation between the number of favorable tweets and outcomes. For BBB, the study presents only the last four evictions, and the researchers observed a correlation between a decrease in positive tweets and the eviction of the Houseguest.

We can find several article reviews on the topic of research. Chauhan, Sharma and Sikka (2020) evaluated 48 articles on sentiment analysis techniques related to predicting election results using data collected from social media. Similarly, Khan *et al.* (2021) employed a systematic review methodology to examine the state of the art. Their study involved a comprehensive search of major computer science digital libraries, including Web of Science, IEEE, ACM, Scopus, and ScienceDirect.

We can organize election forecasting methods into three main approaches to predict election results through social media: Volumetric, Sentiment Analysis (SA), and Social Network Analysis (SNA) (Jaidka *et al.*, 2018; Chauhan; Sharma; Sikka, 2020; Khan *et al.*, 2021; Alvi *et al.*, 2023). The volumetric approach quantifies mentions and interactions, assuming higher values indicate greater support (Skoric *et al.*, 2012; Brito *et al.*, 2019). In sentiment analysis, researchers examine the polarity of sentiments (positive, neutral, and/or negative) in posts to gauge the emotional tone and public opinion. Social network analysis maps the structure and dynamics of interactions to understand opinion dissemination. These methods offer a robust framework for forecasting electoral outcomes based on social media activity

Tumasjan *et al.* (2010) conducted pioneering research by predicting German federal elections based on the number of tweets or publications mentioning a candidate or party. The study collected 70,000 tweets mentioning the six main parties and 35,000 tweets mentioning politicians from those parties. The prediction showed a mean absolute error (MAE) of 1.65%. The volume of messages mentioning a party was a valid indicator of the election outcome, comparable to traditional polls.

Skoric *et al.* (2012) employed the volumetric method to predict the 2011 elections in Singapore. The authors identified and analyzed 110,815 political tweets. They conducted daily and cumulative counts of tweets mentioning various political keywords, including voter names, candidate names, and party/acronym names. The authors based their prediction on the frequency of tweets, resulting in a mean absolute error (MAE) of 5.23%.

In the same way, but using the three main social media platforms, Brito *et al.* (2019) attempted to identify a correlation between candidate performance on social media and votes received at the polls. They conducted a statistical analysis relating votes to the volume of posts, publications and number of followers of the candidates.

In their study of the 2016 Indian state general elections, Sharma and Moh (2016) collected 42,235 Hindi-language tweets concerning five Indian political parties. The researchers employed two distinct sentiment analysis approaches: a supervised method that utilized Naive Bayes and Support Vector Machine (SVM) algorithms, and an unsupervised method that employed a dictionary-based algorithm. The SVM approach demonstrated a 78.4% accuracy rate, correctly predicting that the BJP party would emerge victorious. This prediction was subsequently validated by the party's victory in 60 out of 126 electoral districts.

The goal of Wang and Gan (2017) was to predict the outcome of the 2017 French election. The researchers introduced an equation to measure popularity that included neutral sentiment in the calculation. The inclusion of neutral tweets improved the accuracy of the prediction, demonstrating the importance of considering all sentiments. On the last day before the election, the proposed method predicted a popularity that deviated only about 2% from the actual voting outcome, correctly forecasting Macron's victory.

Several studies have combined sentiment analysis with the volumetric approach to predict election outcomes. Sang and Bos (2012) aimed to predict the 2011 Dutch Senate elections by combining party mentions with sentiment analysis. They collected a total of 64,395 tweets and applied two filters, one party per tweet and one tweet per user, leaving 28,704 tweets. The researchers weighted the number of mentions in this combined approach for more accurate predictions. The authors found that tweet-based predictions came reasonably close to traditional poll results, though some discrepancies remained. This suggests that counting tweets can provide useful election predictions, but data quality and sentiment analysis are critical to improving accuracy.

Bermingham and Smeaton (2011) created a real-time system to analyze the 2011 Irish general election through Twitter. They manually annotated the tweets and employed a supervised approach to classify sentiment using unigrams. Electoral results were most accurately predicted using a linear regression model that incorporated both the volume of mentions and sentiment-based measures as features.

Livne *et al.* (2021) introduced a method employed to predict the results of the 2010 US midterm elections. They constructed a supervised logistic regression model that integrated both structural information from the network and content generated by the candidates. This model successfully determined whether a candidate would win or not.

Ali *et al.* (2022) aimed to use sentiment analysis to predict the outcome of the 2018 general election in Pakistan. They extracted and manually annotated 3,000 tweets. To achieve this, they proposed a machine learning solution, specifically using deep learning.

Chauhan, Sharma and Sikka (2023) employed a lexicon-based sentiment analysis approach to examine Twitter content related to elections and political parties. The study tested various hypotheses by monitoring users' positive sentiments toward different variables throughout the electoral period. The analysis demonstrated that public sentiment on social networks can be a valuable indicator for predicting election outcomes. Furthermore, their findings closely aligned with the actual results of the final election.

This section presented different studies on the prediction of election results. In most cases, the research demonstrates positive outcomes or underscores the potential of utilizing social media for this purpose.

In the present work, we propose a strategy for forecasting in the domain of TV programs such as BBB. The proposal stands out for developing a machine learning model that learns the importance of attributes from previous elections to predict the outcome of future ones.

## 3 Methodology

Considering the potential of social networks and the popularity of the BBB, this experiment seeks to predict the Houseguests who will be eliminated from the "Paredão" based on the texts published by users.

### 3.1 Hypotheses

In this experiment, we have formulated hypotheses and will provide answers based on our findings.

#### 3.1.1 Hypothesis 1: considering all user posts will improve the accuracy of the results

On social media platforms, users can express support or opposition to a Houseguest multiple times. In real-life scenarios like elections, individuals are typically allowed only one vote. However, in Big Brother Brasil, people have the option to submit multiple votes.

#### 3.1.2 Hypothesis 2: The normalization transformation of regression model outputs can significantly influence forecasting outcomes

Regression models often produce outputs that lack an accurate representation of the study's domain; hence, researchers should normalize these outputs.

#### 3.1.3 Hypothesis 3: Applying a spell check during the pre-processing step can enhance the results

Online social networks often feature an internet-specific communication style, including slang and regional terms. Additionally, spelling errors are prevalent. These specifics can impact the results, especially if researchers have not tailored the tool to handle such nuances.

#### 3.1.4 Hypothesis 4: Combining a volumetric strategy with Sentiment Analysis using supervised learning can produce superior results compared to traditional approaches

Studies have demonstrated a range of approaches to predicting election outcomes. Among these, the most popular and successful strategy involves utilizing volumetric analysis combined with sentiment analysis. In the previous section, we presented several relevant pieces of related work that have employed this methodology in forecasting election outcomes.

#### 3.1.5 Hypothesis 5: The proposed solution has the potential to accurately predict the outcome several hours in advance of the official results

Researchers would benefit from considering the potential applications of these proposals to anticipate the results. This could assist professionals in making informed decisions regarding the implementation of changes or the reinforcement of existing strategies.

### 3.2 Data collection

We initially mapped the profiles of each individual on the social network using a list of 22 participants from edition 22 of BBB (Chart 1 presents this information). A recurrent pattern of hashtags emerged throughout the 22nd season of Big Brother Brazil. These hashtags, prefixed with "#Fica" or "#Fora" signified support or opposition, respectively, for particular Houseguest. Users constructed these hashtags by combining the prefix with the name of their chosen participant, sometimes using multiple names for the same individual.

Twitter developed its Application Programming Interface (API) to facilitate the accessibility of Twitter data to both individual users and corporate entities. The academic API provided endpoints for interactions such as replying to comments or searching for tweets. We employed the Python library Tweepy to gather publicly available data from the platform's users and interact with the Twitter academic API. However, Twitter no longer offers the academic API.

In order to collect the tweets from the "Paredão" periods, we used one of the endpoints available in the API, which allowed researchers to obtain data from hashtags over a period of time. Chart 2 delineates the range of dates utilized for data collection, the Houseguest who was evicted, and the targeted outcomes.

For the most part, the showdowns in this edition of the program took place once a week, on Tuesdays. The data collection parameters included the hashtags and the time window. The time window commenced 48 hours prior to the elimination of the participant and concluded at 23:00 on the day of the showdown, as the program typically announced the result after this time.

**Chart 1** - List of BBB 2022 Houseguest and hashtags

| Houseguest | Username | Hashtags |
|---|---|---|
| Luciano Estevan Neto da Conceição | LucianoEstevan | #FicaLuciano, #ForaLuciano |
| Rodrigo Abrão de Carvalho Mussi Ivo | oficialmussi | #FicaRodrigo, #FicaMussi, #FicaRodrigoMussi, #ForaRodrigo, #ForaMussi, #ForaRodrigoMussi |
| Naiara de Fátima Azevedo | Naiarazevedo | #FicaFicaNaiara, #FicaNaiaraAzevedo, #ForaFicaNaiara, #ForaNaiaraAzevedo |
| Vitória Nascimento Câmara (Maria) | eumaria | - |
| Bárbara Heck | bbaheck | #FicaBarbara, #ForaBarbara |
| Bruna Gonçalves Coelho Oliveira (Brunna Gonçalves) | brunnagoncalves | #FicaBrunnaGonçalves, #FicaBrunnaGoncalves, #FicaBrunna, #ForaBrunnaGonçalves, #ForaBrunnaGoncalves, #ForaBrunna |
| Tiago Donato Abravanel Corte Gomes | TiagoAbravanel | - |
| Larissa Tomásia Arruda | larissatomasia | #FicaLarissa, #ForaLarissa |
| Jade Picon Froes | jadepicon | #FicaJade, #FicaJadePicon, #ForaJade, #ForaJadePicon |
| Marcus Vinicius Fernandes de Sousa | vyniof | #FicaViny, #FicaVini, #FicaVinicius, #ForaViny, #ForaVini, #ForaVinicius |
| Laís Rodrigues Caldas | dra_laiscaldass | #FicaLais, #ForaLais |
| Lucas Bissoli Garcia | LucasBissoli_ | #FicaLucas, #ForaLucas |
| Eslovênia Marques de Lima | eslomarques | #FicaEslo, #FicaEslovenia, #ForaEslo, #ForaEslovenia |
| Lina Pereira dos Santos (Linn da Quebrada) | linndaquebrada | #FicaLin, #FicaLinDaQuebrada, #FicaLina, #ForaLin, #ForaLinDaQuebrada, #ForaLina |
| Natália Rocha de Oliveira Deodato | nataliadeodato | #FicaNatalia, #FicaNaty, #ForaNatalia, #ForaNaty |
| Jessilane Alves de Souza | a_jessilane | #FicaJessilane, #FicaJessi, #ForaJessilane, #ForaJessi |
| Gustavo Batista Marsengo | gustavo_beats | #FicaGustavo, #ForaGustavo |

Source: Elaborated by the authors.

The data extracted from the API included the following information:

a) author_id - unique identifier of the user who authored the tweet;
b) tweet_id - unique identifier of the tweet;
c) text - content of the tweet;
d) tweet_conversation_id - identifier of the original tweet of the conversation (including direct replies, replies to replies);
e) tweet_created_at - time of creation of the tweet;
f) tweet_lang - language of the tweet;
g) retweet_count - number of retweets for the respective tweet;
h) reply_count - number of replies to the tweet;

i) like_count - number of likes for the respective tweet;

j) quote_count - number of quotes from the tweet, which is a retweet made with a comment;

k) mentions - mentions of users in the tweet;

l) hashtags - hashtags found in the text.

**Chart 2** - List of BBB 2022 Houseguest and hashtags

| Date of eviction | Houseguests nominated for eviction | Start date | End date | Evicted |
|---|---|---|---|---|
| 1º evicted on January 25th | Luciano (49.31%), Naiara (15.80%), Natália (34.89%) | 23/01/2022 | 25/01/2022 | Luciano |
| 2º evicted on February 1th | Jessilane (25.45%), Natália (26.10%), Rodrigo (48.45%) | 30/01/2022 | 01/02/2022 | Rodrigo |
| 3ª evicted on February 8th | Arthur (3.27%), Douglas (38.96%), Naiara (57.77%) | 06/02/2022 | 08/02/2022 | Naiara |
| 4ª evicted on February 15th | Arthur (4.95%), Bárbara (86.02%), Natália (9.03%) | 13/02/2022 | 15/02/2022 | Bárbara |
| 5ª evicted on February 22th | Brunna (76.18%), Gustavo (22.24%), Paulo André (1.58%) | 20/02/2022 | 22/02/2022 | Brunna |
| 6ª evicted on March 1th | Arthur (5.90%), Larissa (88.59%), Linn (5.51%) | 27/02/2022 | 01/03/2022 | Larissa |
| 7ª evicted on March 8th | Arthur (1.77%), Jade (84.93%), Jessilane (13.30%) | 06/03/2022 | 08/03/2022 | Jade |
| 8º evicted on March 15th | Gustavo (39.51%), Scooby (4.62%), Vinicius (55.87%) | 13/03/2022 | 15/03/2022 | Vinicius |
| 9ª evicted on March 22th | Douglas (4.48%), Eliezer (4.27%), Laís (91.25%) | 20/03/2022 | 22/03/2022 | Laís |
| 10º evicted on March 29th | Lucas (77.54%), Paulo André (4.41%), Scooby (18.05%) | 27/03/2022 | 29/03/2022 | Lucas |
| 11ª evicted on April 3th | Douglas (18.07%), Eslovênia (80.74%), Paulo André (1.19%) | 01/04/2022 | 03/04/2022 | Eslovênia |
| 12ª evicted on April 10th | Eliezer (15.66%), Gustavo (6.74%), Linn (77.60%) | 08/04/2022 | 10/04/2022 | Linn |
| 13ª evicted on April 12th | Gustavo (14.30%), Natália (83.43%), Paulo André (2.27%) | 10/04/2022 | 12/04/2022 | Natália |
| 14ª evicted on April 17th | Arthur (28.54%), Douglas (5.91%), Eliezer (1.92%), Jessilane (63.63%) | 15/04/2022 | 17/04/2022 | Jessilane |
| 15º evicted on April 19th | Eliezer (16.08%), Gustavo (81.53%), Paulo André (2.39%) | 17/04/2022 | 19/04/2022 | Gustavo |

| Date of eviction | Houseguests nominated for eviction | Start date | End date | Evicted |
|---|---|---|---|---|
| 16º evicted on April 21th | Douglas (1.82%), Eliezer (42.23%), Scooby (55.95%) | 19/04/2022 | 21/04/2022 | Scooby |
| 17º evicted on April 24th | Arthur (21.15%), Douglas (13.09%), Eliezer (65.76%) | 22/04/2022 | 24/04/2022 | Eliezer |

Source: Elaborated by the authors.

Table 1 shows the number of tweets extracted from each wall. We collected a total of 1,830,819 tweets, of which 703,433 were unique tweets per user (we selected the most recent tweet from each user). On Twitter, users can repeatedly post the same content or subject, so we constructed two sets: one considering all tweets and the other considering one tweet per user.

**Table 1** - Count of tweets on each eviction process of BBB 2022

| Paredão | All | One tweet per user |
|---|---|---|
| 1º | 53,356 | 33,539 |
| 2º | 158,19 | 67,567 |
| 3º | 135,9 | 62,306 |
| 4º | 103,34 | 54,926 |
| 5º | 101,21 | 45,127 |
| 6º | 53,948 | 28,756 |
| 7º | 435,18 | 174,757 |
| 8º | 51,348 | 25,992 |
| 9º | 93,125 | 33,904 |
| 10º | 55,869 | 20,625 |
| 11º | 20,709 | 7,354 |
| 12º | 72,201 | 21,711 |
| 13º | 28,14 | 12,217 |
| 14º | 24,826 | 54,307 |
| 15º | 81,651 | 21,5 |
| 16º | 93,855 | 24,753 |
| 17º | 44,543 | 14,092 |

Source: Elaborated by the authors.

### 3.3 Pre-processing

Following the acquisition of the data, we preprocessed the texts by removing any numerical data, references, hashtags, and URLs. Furthermore, we conducted a basic spell check, correcting repeated letters only in the accurate number of instances and expanding abbreviated terms when necessary.

Subsequently, we classified the sentiment of each text as positive, negative, or neutral. To perform this sentiment analysis, we employed the vaderSentiment library (Valence Aware Dictionary and Sentiment Reasoner) (Hutto; Gilbert, 2014). This library is based on a lexicon and rules developed for the purpose of analyzing sentiments expressed in social media. Since the library is designed for the English language, we first translated the texts from Portuguese to English using Google Translate before the classification.

### 3.4 Features

For each participant in "Paredão", data was collected on Twitter/X and the following attributes were extracted/calculated:

a) fica count - number of hashtags that have the fica prefix;
b) fora count - number of hashtags that have the fora prefix;
c) fica positive count - number of hashtags that have the prefix fica and for which the sentiment classification of the sentence was given as positive;
d) fica neutral count - number of hashtags that have the prefix fica and for which the sentiment classification of the sentence was given as neutral;
e) fica negative count - number of hashtags that have the prefix fica and for which the sentiment classification of the sentence was given as negative;
f) fora positive count - number of hashtags that have the fora prefix and for which the sentiment classification of the sentence was given as positive;
g) fora neutral count - number of hashtags that have the prefix fora and for which the sentiment classification of the sentence was given as neutral;
h) fora negative count - number of hashtags that have the fora prefix and for which the sentiment classification of the sentence was given as negative;
i) follower rate - number of participant followers divided by the total number of followers of all participants;
j) fica positivity - number of positives fica subtract by number of negatives fica;
k) positivity for a - number of positives fora subtract by number of negatives for a;
l) rate - proportion between hashtags that have the prefix fora and hashtags that have the prefix fica.

Combining the previous attributes, the following sets were created:

a) f0 - fica count, fica neutral count, fica negative count, fica positive count, fora count, fora neutral count, fora negative count, fora positive count;
b) f1 - follower rate, rate, positivity fica, positivity for a;
c) f2 - all features.

### 3.5 Metrics

The following metrics were used to evaluate the performance of the models:

a) mean absolute error (MAE) - is the average of the modules of each error (the difference between the real and the predicted);
b) correct evicted - is the model's percentage of correct answers when indicating the correct participant who was evicted;
c) correct rank - is the percentage of correct answers of the model considering the order in which participants were in the voting result.

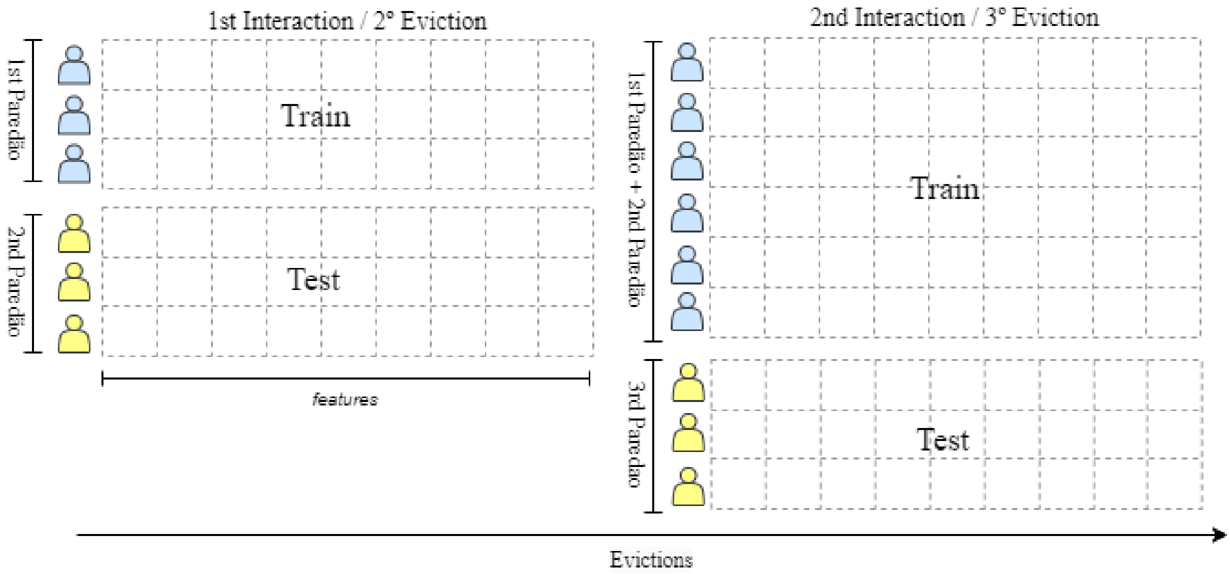
### 3.6 Predicting

We utilized the feature sets to train various models, including Linear Regression, Gradient Boosting Regressor, ElasticNet, Support Vector Regression (with polynomial and linear kernels), Bayesian Ridge Regression, Stochastic Gradient Descent, and Kernel Ridge Regression. We implemented these machine learning algorithms using the Python scikit-learn library.

A TV show's eviction process typically involves three Houseguests, each participant has a set of calculated features. We performed feature calculation using the two sets detailed in Table 2, and utilized selected combinations of these sets for our experiments.

We adopted an iterative approach to assess the performance of the models. In each iteration, we incorporated a new eviction process into the training set and utilized the subsequent eviction process for testing. For instance, in the initial evaluation iteration, we predicted the outcome of the second eviction while training the model on data from the first eviction. Consequently, the test set comprised the target Houseguests in the eviction process, while we constructed the training set using data from preceding evictions. The fundamental assumption of this approach is that researchers should consider each prediction of eviction as a discrete electoral process. In other words, researchers can use historical data on evictions to anticipate future evictions involving different individuals. We evaluated the model's performance based on metrics from the 2nd to the 17th eviction process. Figure 1 illustrates the described methodology.

**Figure 1** - Novel methodology to predict eviction




Source: Elaborated by the authors.

The votes predicted by the regression algorithms may be negative, which is not an acceptable value for the domain in question. Therefore, we normalized the values to a range between 0 and 100 before calculating the MAE. Finally, in each iteration of the forecast, we calculate the mean absolute error (MAE), the number of cases correctly evicted, and the number of cases correctly ranked. We consolidated the ultimate outcomes by computing the mean, standard deviation, and median.

## 4 Results

The preliminary investigation evaluated the impact of the quantity of tweets. In this case, we compare a single tweet per user to all tweets available. Table 2 displays the top-five results of the experiments for the scenarios using all tweets and considering only one tweet per user. The mean MAE results for a single tweet per user indicated a less favorable outcome than those we obtained when utilizing all tweets. Therefore, our results confirm Hypothesis 1. Using all tweets provides more accurate results. This scenario aligns with the domain of BBB, where the audience can vote multiple times.

**Table 2** - Top-5 results of Mean MAE using one tweet per user and all tweets. These results were obtained by training models using a subset of f0 features

| Models | One | All | Improve |
|---|---|---|---|
| RandomForestRegressor | 20.481 | 16.560 | 23.68% |
| BayesianRidge | 18.086 | 17.993 | 0.51% |
| SGDRegressor | 18.877 | 18.674 | 1.09% |
| RandomForestRegressor | 19.022 | 18.689 | 1.78% |
| GradientBoostingRegressor | 19.979 | 19.229 | 3.90% |

Source: Elaborated by the authors.

In the subsequent investigation, we transformed the model results into domain targets using all tweets. Table 3 presents the impact of normalization on improved results, which supports Hypothesis 2.–The transformation to domain targets indicates that this approach enhances the results.

**Table 3** - Top-five results of Mean MAE before and after normalization. These results were obtained by training models using a subset of f0 features and all of the tweets

| Models | Before | After | Improve |
|---|---|---|---|
| ElasticNet | 26.632 | 11.242 | 136.89% |
| GradientBoostingRegressor | 20.287 | 12.038 | 68.53% |
| KernelRidge | 29.169 | 12.659 | 130.42% |
| SVR_Linear | 25.033 | 12.749 | 96.35% |
| SGDRegressor | 25.717 | 13.456 | 91.11% |

Source: Elaborated by the authors.

To enhance the outcomes of our preliminary investigations that examine the influence of implementing a simple spell-checking process prior to the prediction, Table 4 presents the impact of the spell-checking process on sentiment classification. We observed that a small number of instances, representing less than one percent of the total data set, exhibited a change in polarity from neutral to either positive or negative following the spell-checking process.

**Table 4** - Classification of polarity sentiment: a comparison of the percentage of tweets before and after spell check

| Sentiment | Without Spell Check | After Spell Check |
|---|---|---|
| Fica Neutral | 56.44% | 55.64% |
| Fica Negative | 16.78% | 16.92% |
| Fica Positive | 26.78% | 27.44% |
| Fora Neutral | 58.80% | 57.94% |
| Fora Negative | 16.96% | 17.13% |
| Fora Positive | 24.24% | 24.93% |

Source: Elaborated by the authors.

In the scenario where we used all the characteristics and all the tweets, the top-five best results show a slight improvement in most cases (Table 5). The incorporation of spell-checking into this methodology is potentially good for obtaining improved results. However, there is room for improvement in the current spell-checking process.

**Table 5** - Top-five results of Mean MAE before and after Spell Check. These results were obtained by training models using a subset of f0 features and all of the tweets

| Models | Before | After | Improve |
|---|---|---|---|
| ElasticNet | 11.254 | 11.242 | 0.10% |
| GradientBoostingRegressor | 12.016 | 12.038 | -0.18% |
| KernelRidge | 12.656 | 12.659 | -0.03% |
| SVR_Linear | 12.775 | 12.749 | 0.20% |
| SGDRegressor | 13.504 | 13.456 | 0.35% |
| BayesianRidge | 14.384 | 14.301 | 0.58% |

Source: Elaborated by the authors.

**Table 6** - The results of the mean MAE, mean correct eviction, and mean Correct rank. These results were obtained by training models using different subsets of features and all the tweets

| Features | Models | Mean MAE | Mean Correct Evicted | Mean Correct rank |
|---|---|---|---|---|
| f0 | LinearRegression | 14.761 | 75.00% | 56.25% |
| f0 | GradientBoostingRegressor | 12.038 | **81.25%** | 56.25% |
| f0 | ElasticNet | **11.242** | **81.25%** | **68.75%** |
| f0 | SVR_Poly | 20.085 | 56.25% | 25.00% |
| f0 | SVR_Linear | 12.749 | 75.00% | 62.50% |
| f0 | BayesianRidge | 14.301 | 68.75% | 56.25% |
| f0 | SGDRegressor | 13.456 | 68.75% | 56.25% |
| f0 | KernelRidge | 12.659 | **81.25%** | 62.50% |
| f0 | DecisionTreeRegressor | 17.279 | 68.75% | 50.00% |
| f0 | RandomForestRegressor | 14.646 | 75.00% | 50.00% |
| f1 | LinearRegression | 17.689 | 62.50% | 43.75% |
| f1 | GradientBoostingRegressor | 21.375 | 50.00% | 43.75% |
| f1 | ElasticNet | 18.156 | 56.25% | 56.25% |
| f1 | SVR_Poly | 23.645 | 62.50% | 50.00% |
| f1 | SVR_Linear | 19.394 | 62.50% | 56.25% |
| f1 | BayesianRidge | 19.167 | 62.50% | 50.00% |
| f1 | SGDRegressor | 30.968 | 37.50% | 25.00% |
| f1 | KernelRidge | 17.687 | **81.25%** | 43.75% |
| f1 | DecisionTreeRegressor | 16.641 | 62.50% | 50.00% |
| f1 | RandomForestRegressor | 18.260 | 50.00% | 50.00% |
| f2 | LinearRegression | 18.947 | 62.50% | 43.75% |
| f2 | GradientBoostingRegressor | 17.145 | 62.50% | 50.00% |
| f2 | ElasticNet | 17.625 | 62.50% | 62.50% |
| f2 | SVR_Poly | 23.618 | 62.50% | 50.00% |
| f2 | SVR_Linear | 18.673 | 56.25% | 56.25% |
| f2 | BayesianRidge | 19.061 | 56.25% | 50.00% |
| f2 | SGDRegressor | 29.208 | 43.75% | 31.25% |
| f2 | KernelRidge | 14.390 | **81.25%** | 56.25% |
| f2 | DecisionTreeRegressor | 16.775 | 68.75% | 56.25% |
| f2 | RandomForestRegressor | 14.892 | 68.75% | 56.25% |

Source: Elaborated by the authors.

**Table 7** - Baseline methods of forecasting

| Models | Mean MAE | Mean Correct evicted | Mean Correct rank |
|---|---|---|---|
| Fora | **9.031** | 87.50% | 62.50% |
| Positive Fora | 9.042 | 87.50% | 62.50% |
| Fora (-8h) | 9.295 | 87.50% | 62.50% |
| Positive Fora (-8h) | 9.283 | 87.50% | 62.50% |

Source: Elaborated by the authors.

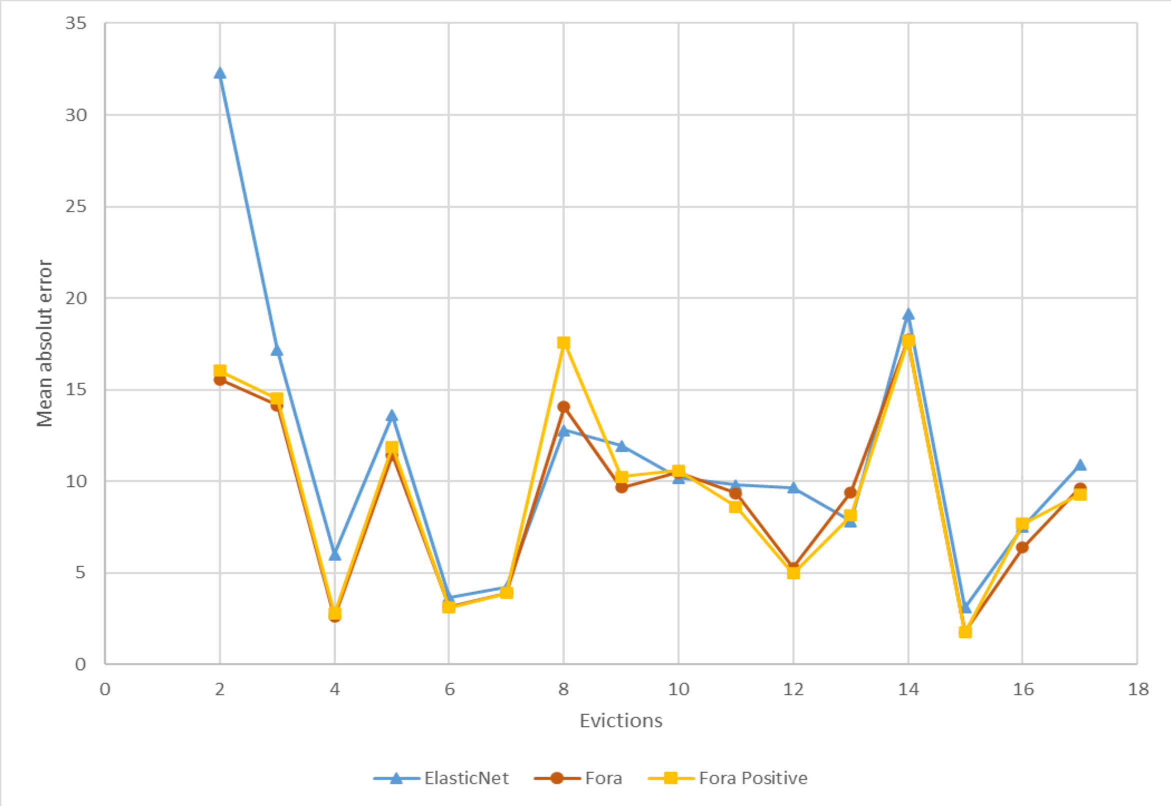
In accordance with the preceding Hypothesis verification, we conducted an experiment with various models utilizing default hyperparameters and subsets of features, as detailed in Subsection 3.4. Table 6 presents the results of this experiment. The ElasticNet training with features that combine volumetric and sentiment analysis resulted in superior outcomes in the experiments. The mean MAE was 11.42, the mean correct eviction was 81.25%, and the mean correct rank was 68.75%.

Table 7 presents the findings of two most used methods as a benchmark in this study. The baseline methods include the counting of hashtags with Fora and the counting of sentiments with Fora positive, including the case with eight hours less time. The results do not show a significant difference. Although the best MAE results indicate proximity, the confirmation of Hypothesis four remains uncertain.

In terms of average MAE measurement, the Fora volumetric method achieved the best result, with an average of 9.031. In addition, the baseline methods achieved the best result in terms of the average correct eviction rate of 87.50%. However, the ElasticNet model with the f0 feature subset achieved the highest mean correct rank result (68.75%).

Figure 2 provides a detailed illustration of the performance in metric MAE. It shows that the primary discrepancy occurs during the initial eviction. The limited historical data available for model training causes this inconsistency.

**Figure 2** - Compare the best model and baselines methods for each eviction




Source: Elaborated by the authors.

**Table 8** - The results of the mean MAE, mean correct eviction, and mean correct rank. These results were obtained in prediction 8 hours before

| Features | Models | Mean MAE | Mean Correct Evicted | Mean Correct rank |
|---|---|---|---|---|
| f0 | LinearRegression | 13.729 | 75.00% | 56.25% |
| f0 | GradientBoostingRegressor | 12.142 | 68.75% | 56.25% |
| f0 | ElasticNet | **11.847** | **81.25%** | **68.75%** |
| f0 | SVR_Poly | 20.511 | 56.25% | 25.00% |
| f0 | SVR_Linear | 12.958 | 81.25% | 68.75% |
| f0 | BayesianRidge | 13.920 | 75.00% | 62.50% |
| f0 | SGDRegressor | 13.006 | 75.00% | 62.50% |
| f0 | KernelRidge | 12.948 | 81.25% | 62.50% |
| f0 | DecisionTreeRegressor | 13.845 | 75.00% | 68.75% |
| f0 | RandomForestRegressor | 13.776 | 81.25% | 56.25% |
| f1 | LinearRegression | 16.921 | 62.50% | 50.00% |
| f1 | GradientBoostingRegressor | 19.894 | 50.00% | 43.75% |
| f1 | ElasticNet | 18.120 | 56.25% | 56.25% |
| f1 | SVR_Poly | 24.264 | 62.50% | 50.00% |
| f1 | SVR_Linear | 19.495 | 56.25% | 50.00% |

| | | | | |
|---|---|---|---|---|
| f1 | BayesianRidge | 18.999 | 62.50% | 50.00% |
| f1 | SGDRegressor | 30.523 | 37.50% | 25.00% |
| f1 | KernelRidge | 17.772 | 81.25% | 50.00% |
| f1 | DecisionTreeRegressor | 19.022 | 56.25% | 43.75% |
| f1 | RandomForestRegressor | 18.357 | 56.25% | 50.00% |
| f2 | LinearRegression | 18.797 | 62.50% | 43.75% |
| f2 | GradientBoostingRegressor | 18.820 | 50.00% | 37.50% |
| f2 | ElasticNet | 17.809 | 62.50% | 62.50% |
| f2 | SVR_Poly | 24.234 | 62.50% | 50.00% |
| f2 | SVR_Linear | 18.725 | 56.25% | 56.25% |
| f2 | BayesianRidge | 18.873 | 56.25% | 50.00% |
| f2 | SGDRegressor | 29.004 | 43.75% | 31.25% |
| f2 | KernelRidge | 14.607 | 81.25% | 56.25% |
| f2 | DecisionTreeRegressor | 15.772 | 75.00% | 56.25% |
| f2 | RandomForestRegressor | 14.302 | 68.75% | 62.50% |

Source: Elaborated by the authors.

In this experiment, we considered a reduction of eight hours in the test data, with Table 8 displaying these results. The optimal model we identified in this experiment remained consistent when we considered all data (Table 6), exhibiting a difference of only 0.605 in the MAE metric and obtaining the same results in the other metrics. Although the results differed from the target value, they were similar to the best result we obtained using 48-hour data. This suggests that the solution has the potential to predict outcomes, particularly in terms of accurate eviction predictions and the identification of the correct eviction sequence.

## 5 Discussion

In contrast to other studies, we employed a supervised strategy that integrated sentiment, tweet volume, and positioning as inputs for our models. When we examined the sentiment analysis results presented in Table 7, we found that the distinct contextual differences in our experimental setup prevented us from drawing positive conclusions that align with the findings of Bermingham and Smeaton (2011), Sang and Bos (2012), Wang and Gan (2017), and Chauhan, Sharma and Sikka (2023). This observation challenges the potential for direct comparison with the simpler volumetric approach that Skoric *et al.* (2012) and

Tumasjan *et al.* (2010) used in their methodology. Nevertheless, researchers may consider alternative counting metrics as predictors, as evidenced by Brito *et al.* (2019), who identified a positive correlation with the number of followers.

One potential avenue for sentiment analysis that may demonstrate favorable outcomes is the utilization of a supervised methodology, which contrasts with the unsupervised lexicon approach that we employed in this study. Ali *et al* (2022) employed deep learning for this purpose but identified a limiting factor in the form of data quantity. Furthermore, researchers must label the data, ideally within the context of the intended application, which can also act as a limiting factor. Sang and Bos (2012) highlight the importance of data selection, a point that we also addressed in our experiments when we specified one tweet per user versus using all tweets. This approach proved crucial for analyzing the impact of data filtering.

## 6 Conclusion

We devised a methodology to integrate disparate approaches and demonstrate their application to a practical problem: forecasting popular voting outcomes in a complex scenario involving sequential processes of voting. The results suggest the potential for predicting the outcomes of television shows and elections using social media data.

Furthermore, the traditional approach exhibited superior performance in most metrics compared to the machine learning approach, particularly in terms of sentiment and volumetric analysis. We particularly highlight that the proposed method demonstrated superior accuracy in predicting the correct rank.

Based on the experimental results, we identified future directions for this project. To enhance the machine learning approach, we propose several avenues of exploration, including the optimization of algorithm hyperparameters, the incorporation of new features, and the enhancement of pre-processing and normalization techniques. Another potential direction would be to employ alternative sentiment analysis methods, add user positioning information, and incorporate data from social network analysis. Additionally, leveraging new data sources can provide a more comprehensive understanding, as these sources may introduce novel biases.

## References

ALI, Haider; FARMAN, Haleem; YAR, Hikmat; *et al*. Deep learning-based election results prediction using Twitter activity. **Soft Computing**, New York, v. 26, n. 16, p. 7535-7543, 2022. Available at: https://doi.org/10.1007/s00500-021-06569-5 . Access: 26 Aug. 2024.

ALVI, Quratulain; ALI, Syed Farooq; AHMED, Sheikh Bilal; *et al*. On the frontiers of Twitter data and sentiment analysis in election prediction: a review. **PeerJ Computer Science**, Liverpool, v. 9, p. 1-25, 2023. Available at: https://doi.org/10.7717/peerj-cs.1517 . Access: 26 Aug. 2024.

BERMINGHAM, Adam; SMEATON, Alan. On using Twitter to monitor political sentiment and predict election results. *In*: WORKSHOP ON SENTIMENT ANALYSIS WHERE AI MEETS PSYCHOLOGY, 2011, Chiang Mai, Thailand. **Proceedings** […]. [*S. l.*]: Asian Federation of Natural Language Processing, 2011. p. 2-10.

BRITO, Kellyton Santos; SILVA FILHO, Rogerio Luiz Cardoso; ADEODATO, Paulo Jorge Leitao. A systematic review of predicting elections based on social media data: research challenges and future directions. **IEEE Transactions on Computational Social Systems**, New York, v. 8, n. 4, p. 819-843, 2021. Available at: https://doi.org/10.1109/TCSS.2021.3063660 . Access: 26 Aug. 2024.

BRITO, Kellyton; PAULA, Natalia; FERNANDES, Manoel; *et al*. Social media and presidential campaigns - preliminary results of the 2018 Brazilian presidential election. *In*: ANNUAL INTERNATIONAL CONFERENCE ON DIGITAL GOVERNMENT RESEARCH, 20., 2019, Dubai. **Proceedings** […]. New York: ACM, 2019. p. 332-341.

CHAUHAN, Priyavrat; SHARMA, Nonita; SIKKA, Geeta. Application of Twitter sentiment analysis in election prediction: a case study of 2019 Indian general election. **Social Network Analysis and Mining**, New York, v. 13, n. 88, p. 1-29, 2023. Available at: https://doi.org/10.1007/s13278-023-01087-8 . Access: 26 Aug. 2024.

CHAUHAN, Priyavrat; SHARMA, Nonita; SIKKA, Geeta. The emergence of social media data and sentiment analysis in election prediction. **Journal of Ambient Intelligence and Humanized Computing**, New York, v. 12, n. 2, p. 2601-2627, 2020. Available at: https://doi.org/10.1007/s12652-020-02423-y . Access: 26 Aug. 2024.

FIGUEREDO, Igleson; MARINHO, Leandro; ALVES, Leonardo. Investigando a influência de tweets em programas de votação popular no Brasil. *In*: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING, 2015. Recife. **Anais** […]. Porto Alegre: Sociedade Brasileira de Computação, 2015.

HUTTO, C.; GILBERT, Eric. VADER: A Parsimonious rule-based model for sentiment analysis of social media text. **Proceedings of the International AAAI Conference on Web and Social Media**, Washington, v. 8, n. 1, p. 216-225, 2014. Available at: https://doi.org/10.1609/icwsm.v8i1.14550 . Access: 26 Aug. 2024.

JAIDKA, Kokil; AHMED, Saifuddin; SKORIC, Marko; *et al*. Predicting elections from social media: a three-country, three-method comparative study. **Asian Journal of Communication**, London, v. 29, n. 3, p. 252-273, 2018. Available at: https://doi.org/10.1080/01292986.2018.1453849 . Access: 26 Aug. 2024.

KHAN, Asif; ZHANG, Huaping; BOUDJELLAL, Nada; *et al*. Election prediction on twitter: a systematic mapping study. **Complexity**, New Jersey, v. 2021, n. 5, p. 1-27, 2021. Available at: https://doi.org/10.1155/2021/5565434 . Access: 26 Aug. 2024.

LIVNE, Avishay; SIMMONS, Matthew; ADAR, Eytan; *et al*. The party is over here: structure and content in the 2010 election. *In*: INTERNATIONAL AAAI CONFERENCE ON WEBLOGS AND SOCIAL MEDIA, 5., 2021, Barcelona. **Proceedings** […].Washington: Association for the Advancement of Artificial Intelligence, 2021. v. 5, p. 201-208.

ROUSIDIS, Dimitrios; KOUKARAS, Paraskevas; TJORTJIS, Christos. Social media prediction: a literature review. **Multimedia Tools and Applications**, New York, v. 79, n. 9-10, p. 6279-6311, 2019. Available at: https://doi.org/10.1007/s11042-019-08291-9 . Access: 26 Aug. 2024.

SANG, Erik Tjong Kim; BOS, Johan. Predicting the 2011 dutch senate election results with Twitter. *In*: WORKSHOP ON SEMANTIC ANALYSIS IN SOCIAL MEDIA, 2012, Avignon. **Proceedings** […].Kerrville: Association for Computational Linguistics, 2012. p. 53-60.

SHARMA, Parul; MOH, Teng-Sheng. Prediction of Indian election using sentiment analysis on Hindi Twitter. *In*: IEEE INTERNATIONAL CONFERENCE ON BIG DATA, 2016, Washington. **Proceedings** […]. New York: IEEE, 2016. p. 1966-1971.

SKORIC, Marko; POOR, Nathaniel; ACHANANUPARP, Palakorn*; et al*. Tweets and votes: a study of the 2011 Singapore general election. *In*:HAWAII

INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES, 45., 2012, Maui. **Proceedings** […]. New York: IEEE, 2012. p. 2583-2591.

SOLER, J. M.; CUARTERO, F.; ROBLIZO, M. Twitter as a tool for predicting elections results. *In*: IEEE/ACM INTERNATIONAL CONFERENCE ON ADVANCES IN SOCIAL NETWORKS ANALYSIS AND MINING, 2012, Istanbul. **Proceedings** […]. New York: IEEE, 2012. p. 1194-1200.

TUMASJAN, Andranik; SPRENGER, Timm; SANDNER, Philipp; *et al*. Predicting elections with twitter: what 140 characters reveal about political sentiment. *In*: INTERNATIONAL AAAI CONFERENCE ON WEBLOGS AND SOCIAL MEDIA, 4., 2010. **Proceedings** […].Washington: AAAI, 2010. p. 178-185.

WANG, Lei; GAN, John Q. Prediction of the 2017 French election based on Twitter data analysis. *In*: COMPUTER SCIENCE AND ELECTRONIC ENGINEERING, 9., 2017, Colchester. **Proceedings** […]. New York: IEEE, 2017. p. 89-93.

## Quem será eliminado do Big Brother Brasil? Previsão baseada em redes sociais

**Resumo:** O advento das mídias sociais forneceu aos indivíduos um ambiente para a expressão de opiniões e preferências em uma ampla gama de assuntos. Os dados obtidos das mídias sociais podem ser uma fonte valiosa de informações para a análise das intenções e interesses do público, facilitando assim o processo de tomada de decisão e desenvolvimento de estratégias. Um exemplo de análise do comportamento humano é o exame de eventos de votação popular. Para abordar o desafio de prever uma sequência de eventos no contexto da votação, propomos uma nova metodologia que emprega uma solução orientada por dados, incorporando dados do Twitter/X e modelos de regressão. Neste estudo de caso, empregamos técnicas comumente utilizadas na previsão de resultados eleitorais, incluindo análise volumétrica e de sentimentos, para prever a eliminação de concorrentes no Big Brother Brasil. Nossos experimentos resultaram em um erro absoluto médio de aproximadamente 11 pontos, com uma precisão de 81,25% para prever eliminações e 68,75% para prever a ordem da classificação durante cada eliminação.

**Palavras-chave:** previsão; Big Brother Brasil; análise de sentimento; análise volumétrica; análise de redes sociais

**Authorship and Responsibility Statement**

**Conception and design of the study:** William Takahiro Maruyama, Luciano Antonio Digiampietri
**Data collection:** William Takahiro Maruyama, Luciano Antonio Digiampietri
**Data analysis and interpretation:** William Takahiro Maruyama, Luciano Antonio Digiampietri
**Writing:** William Takahiro Maruyama, Luciano Antonio Digiampietri
**Critical review of the manuscript:**William Takahiro Maruyama, Luciano Antonio Digiampietri

**Data availability statement**

The dataset supporting the results of this study can be requested from the authors (see corresponding author).

**Corresponding authorship**

William Takahiro Maruyama

williammaruyama@usp.br

**Editor-in-chief**

Thiago Henrique Bragato Barros

**How to cite**
MARUYAMA, William Takahiro; DIGIAMPIETRI, Luciano Antonio. Who will be evicted from Big Brother Brazil? Forecasting based on social networking. **Em Questão**, Porto Alegre, v. 31, e-142154, 2025. https://doi.org/10.1590/1808-5245.31.142154.

**Open peer review:**

https://doi.org/10.1590/1808-5245.31.142154A

Receveid: 28/08/2024
Accepted: 28/04/2025