Predicting Engagement of Brazilian Politicians on TikTok: A Machine Learning Approach

Maria Santana[1], José Santana[2], Pablo Sampaio[1(✉)], and Kellyton Brito[1]

[1] Departamento de Computação, Universidade Federal Rural de Pernambuco, Recife, Brazil
{mariagabrielly.santana,pablo.sampaio,kellyton.brito}@ufrpe.br
[2] DEINFO, Universidade Federal Rural de Pernambuco, Recife, Brazil
josecarlos.santana@ufrpe.br

Abstract. While established social media platforms like Facebook, Twitter, and Instagram have become staples in political campaigns, the 2022 Brazilian elections witnessed the rise of a new contender: TikTok. Despite its recent emergence in 2016, TikTok has already become the fourth most used social network in Brazil. This study investigates the potential of machine learning to predict engagement on the TikTok profiles of the two leading presidential candidates: Lula and Bolsonaro. Utilizing a dataset from previous studies, we implemented various machine learning models and found that the Support Vector Machine achieved the highest performance based on the F1-score metric for both candidates. Despite the results being better with Bolsonaro than with Lula, further analysis of metrics like recall and precision suggests valuable insights for social and political domains. These findings can aid both candidates and society in understanding what factors are most related to engagement on this emerging social media platform. Additionally, marketing and advertising teams can use this information to create content tailored to reach and engage with a politician's target electorate.

Keywords: TikTok · Political Campaigns · Social Media · Engagement Prediction · Machine Learning · Brazilian Election

1 Introduction

Despite being relatively new, social media platforms are now an integral part of daily life for millions of Brazilians. WhatsApp, Instagram and Facebook are the three more accessed platforms by Brazilian users. Facebook Messenger, Telegram, Pinterest, Kuaishou (Kwai), Twitter and LinkedIn follow as the next most popular platforms [1].

Within this context, Brazilian politicians are active users of these platforms, employing them as communication tools to engage with their constituents and conduct political campaigns. Recent Brazilian elections have demonstrated this trend, such as in 2018 when former President Bolsonaro was elected with minimal exposure on broadcast television. His electoral success is often attributed to his presence on social media [2, 3].

Previous research has attempted to identify correlations between elements of politicians' posts and profiles and electoral outcomes. While some studies have yielded results

A. Paes and F. A. N. Verri (Eds.): BRACIS 2024, LNAI 15414, pp. 387–401, 2025.
https://doi.org/10.1007/978-3-031-79035-5_27

that support this hypothesis [4], others lack conclusive findings and are still undergoing testing of their approaches [5].

Among the most popular social media platforms in Brazil today, a new platform aimed at young audiences [6] stands out: TikTok. This platform was largely ignored by candidates in the 2018 Brazilian presidential election but was heavily utilized in the 2022 election. However, despite being the sixth most used social network in the world [7], due to its early stage of usage in electoral campaigns, it is still challenging to find studies that analyze and predict outcomes within the context of elections. An exception presented in [8] is an analysis of the content of videos on this platform by the main presidential candidates in 2022, Lula and Bolsonaro, based on creating and applying a taxonomy to identify correlations between the type of content and candidate engagement. However, the dataset was not used for predicting engagement.

In this context, this article presents the results of a study that aims to predict engagement on TikTok based on the use of the platform by the two main candidates in the 2022 Brazilian presidential election. Thus, it aims to answer two research questions: RQ1: To what extent can discourse features effectively predict engagement levels in political posts using basic Machine Learning models? And RQ2: Which discourse features are the most important for predicting the number of likes (engagement) based on the trained Machine Learning models? For this, we used the classification taxonomy and the data collected and classified in [8] to train several models for the first time, including neural networks, support vector machine, random forest, k-nearest neighbors, and logistic regression.

The remainder of this paper is as follows. In Sect. 2, related works will be addressed. Section 3 discusses the methodology used in this study. Section 4 presents the results. Section 5 covers the discussions regarding the results, while Sect. 6 presents conclusions and future work.

2 Related Works

Starting with Barack Obama's 2008 campaign [9], political strategy has been revolutionized over the last 15 years through social media, sparking researchers' interest and numerous studies on social medias' impact on political campaigns.

Notable studies include [10] investigation of Donald Trump's 2016 campaign in the United States and [3] study of Jair Bolsonaro's 2018 campaign in Brazil. More specifically, some research seeks to identify a possible correlation between the engagement generated on candidates' social media platforms and the number of votes they receive in elections.

The work of [11] was groundbreaking in proposing that the number of tweets mentioning a political party could serve as a reliable indicator of the percentage of votes it would receive, demonstrating predictive power like traditional opinion polls. Most subsequent studies followed this line of thought, focusing on Twitter analysis [4]. However, this Twitter-centric approach has faced increasing criticism [12]. One of the main objections points out that Twitter is not among the top 10 most used social networks globally, raising questions about the representativeness of the analyzed data and several biases related to data gathering.

Working from a sentiment analysis perspective, a recent work [13] proposed a model for election prediction using sentiment analysis of influential messages on Facebook in the 2016 United States presidential election. The results demonstrated better classification performance with the Random Forest algorithm and that predicting election results based on sentiment analysis of messages was reliable.

With a multi-platform approach, Vepsäläinen and Suomi [14] investigates the role of two different social media platforms, Twitter and Facebook, in predicting the 2019 Finnish parliamentary elections. The results demonstrated that both the number of likes on Facebook and the number of followers on Twitter are positively associated with election results. The test results suggest that the number of followers on Twitter and likes on Facebook are more influential in determining the election outcome for candidates with less political experience.

It is clear, therefore, that the platform most used by researchers for political data analysis is Twitter, with most applied methods focusing on volume and sentiment analysis of mentions. This approach is widely criticized [15]. Among the reasons is the fact that in 2022, Twitter was the fifteenth most-used platform in the world [16] and the ninth in Brazil [7]. Meanwhile, TikTok is the sixth most used platform in the world [16], the fourth most used in Brazil [7], and preferred by young people [6]. As a recent network, politicians are still starting to use TikTok, and there is little research focused on analysis and prediction regarding this social media and electoral results.

Thus, the present work uses features based on a deep analysis of the discourse found in videos on the TikTok platform found in [8] using the proposed taxonomy and the dataset used in that same research. Most of the other works are less based on content, but this one attempts to make predictions based on a deep analysis of the content to assess the predictive capacity of this analysis.

3 Methodology

This study aims to investigate the relationship between discourse features and engagement levels in political posts on social media. Specifically, we seek to answer the following research questions, segmented per candidate:

- *RQ1*: To what extent can discourse features effectively predict engagement levels in political posts using basic Machine Learning models?
- *RQ2*: Which discourse features are the most important for predicting the number of likes (engagement) based on the trained Machine Learning models?

To answer these research questions, we defined a methodology consisting of five steps: data collection, preprocessing, data transformation, model training, and analysis of the training results.

3.1 Data Collection

The data utilized in this study were obtained from a public dataset used in previous works [8, 17, 18]. While this dataset has been employed in various data analyses before, this is the first study to use it for developing predictive machine learning models. The dataset

includes features from videos posted by the two main candidates in the 2022 Brazilian presidential election, Lula and Bolsonaro, on the TikTok platform.

The data collection period spanned from June 30th to October 30th, totaling 122 days of collection. The studies focused on the two main candidates for the presidency of Brazil, Lula (@lulaoficial) and Bolsonaro (@bolsonaromessiasjair), which together received 92% of the total votes in the first round. The data was manually classified with attributes related to features of the discourse of the candidates observed in the videos (TikTok posts), based on a predefined taxonomy [8]. More details of the data collected and their features are provided in Sect. 4.1.

3.2 Preprocessing

The preprocessing involves: data cleaning, manual feature engineering, data subdivision and target attribute specification. Regarding *data cleaning*, we simply removed posts with missing values in any of its attributes.

During preprocessing, we *manually engineered* a new feature from the original dataset that may be relevant to the classification task. The feature, called "Days Elapsed," represents the number of days between the publication date of a post and the date the data was collected for that post (to form the dataset). This can be justified by the fact that the posts in our dataset were not exposed for proportionally similar periods and that they were released in different periods of the campaign. Regarding *data subdivision*, we *separated* the datasets into Bolsonaro Posts and Lula Posts to enable independent analysis for each political figure.

A special care was taken regarding the *target attribute specification.* The basic raw attribute used to measure engagement in this work is the number of likes received by a post. This metric offers several advantages: it is unique to each user, preventing duplication of engagement measures, and it exhibits a linear correlation with other potential engagement metrics such as the number of comments, shares, and plays. Then, we opted for a classification approach. We divided the posts of each candidate into two distinct classes based on their engagement levels: high engagement, for posts with engagement levels at or above the median (50th percentile) of a candidate's posts; and low engagement, encompassing the other half of the posts.

We chose to treat this study as a classification problem rather than a regression problem because we understand that it would have similar practical relevance in a system designed to recommend whether or not to proceed with a post (before actually posting), for example. Additionally, creating models that accurately capture the distribution of likes can be particularly challenging due to the highly non-linear nature of engagement (for instance, small differences in content can lead to a post going viral or not).

3.3 Data Transformation

We have tested different transformations for categorical features and numerical features. After applying the transformations of the categorical features, the resulting features are also subject to the second type of transformations.

Categorical Features. Since most of the features in the data are categorical, it is especially important to evaluate the impact of different transformations for this type of feature. Therefore, we evaluated three alternate transformation approaches:

- *One-hot encoding*: converts a categorical feature with N categories into N binary features, each indicating the presence of a specific category.
- *Target encoding for (binary) classification*: transforms categorical variables into continuous variables based on their relationship with the target classes in the training data [19].
- *Target encoding for regression*: a version of target encoding for continuous target variables [19]. We hypothesize that applying the target encoding to the raw values of likes could render useful information for the binary classification task.

Numerical Features. For machine learning models that are sensitive to the scale of the data, we also tested three alternative transformations for numerical features.

- *Min-Max (0–1) Scaling*: This technique sets the minimum value of each feature to 0 and the maximum value to 1, linearly scaling all other values proportionally within this range.
- *Standard Scaling* (Z-score normalization): Standardize the data to have a mean of 0 and a standard deviation of 1.
- *Identity*: No transformation is applied; the original values are used.

To ensure the integrity of our model and prevent data leakage, both types of transformation were applied strictly to the training data within the training pipeline (and not to the full dataset, as a preprocessing step). One-hot encoding, however, is always applied considering all the categories of the full dataset, since these were part of a pre-defined taxonomy [8].

3.4 Model Training

Different classification models were evaluated with different hyperparameters in this study, as detailed in Table 1. Hyperparameters not listed received the default values of the scikit-learn 1.4 implementation (e.g., MLP used the ADAM optimizer), except for logistic regression, where SAGA solver was chosen for its flexibility regarding the penalty term. For model type, a grid search approach was employed to evaluate all combination of feature transformations and hyperparameters. In this context, the transformations were treated similarly to hyperparameters to explore their impact on the model performance. We use the term model configuration to refer to the hyperparameters of the model combined with the transformations to be applied to the input data.

Nested Cross-Validation Procedure. To robustly train and evaluate each type of model given the relatively limited amount of data, we employed a nested 5-fold cross-validation (CV) approach. This technique provides a more reliable estimate of model performance and helps mitigate the risk of overfitting. The 5-fold nested CV procedure involves splitting the data into 5 folds and iterating five times. In each iteration, 4 folds are used

Table 1. Models and set of hyperparameter values evaluated.

Model Type	Hyperparameters and Values
MLP Neural Network (with 1 hidden layer) (MLP)	hidden layer size: 16, 64 or 256 activation function: relu, tanh or logistic learning rate: 0.001, 0.01, 0.1
Support Vector Machine (SVM)	C: 0.1, 1.0, 10.0, 20.0 or 50.0 gamma: 'scale' or 'auto' kernel: rbf or sigmoid class weight: 'balanced' or none
Random Forest (RF)	number of trees: 10, 30 or 70 maximum depth: 3, 4, 5 or free minimum samples before splitting: 2, 4 or 8 class weight: 'balanced' or none
Logistic Regression (LR)	C: 0.01, 0.1, 1.0 or 2.0 penalty term: L1, L2 or none class weight: 'balanced' or none
K-Nearest Neighbors (KNN)	number of neighbors: 5, 10, 15 or 20 weights: uniform or distance-based p: 1 or 2

for training, while the remaining fold is held out as a test set for evaluation. Each fold maintained an equal class distribution, with 50% for each class. In each iteration of the outer CV, the grid search is applied to this training (4 folds of data) to find the optimal *model configuration* for that iteration. In each run of the grid search, each model configuration (i.e. combination of hyperparameters and transformations) was evaluated using an inner 5-fold CV applied to the training data. This inner CV further splits the training data into 4 inner folds for model fitting and 1 inner fold for validation. The average F1-score of the 5 iterations of this inner CV is used to assess each model configuration. As a result of each run of grid search (in each iteration of the outer CV), one *optimal model configuration* is identified for each iteration of the outer CV. So, in total, five optimal configurations will be found per model type.

Model Evaluation. At the end of each iteration of the outer CV, the model is retrained using the optimal configuration on the entire training data (i.e. the four selected outer folds). The model is then evaluated on the held-out test fold using these metrics: accuracy, precision, recall, and F1-score.

We highlight that one optimal model configuration (hyperparameters + transformations) is obtained and evaluated per iteration of the outer CV, thus resulting in five configurations as well as five values for each of the evaluation metrics, per model type. These values and settings are used as inputs for the subsequent analysis step.

3.5 Analysis of Results

For each of research questions proposed we performed a specific analysis, as detailed next.

Model Performance Analysis. To address RQ1, we first calculated and reported the mean and standard deviation of the evaluation metrics for each model type and dataset. To further understand how to effectively predict the engagement, we analyzed the optimal *model configurations* obtained, reporting the most frequent data transformations chosen in the grid search. We also discuss key differences or similarities observed in model performance and configuration between Bolsonaro's and Lula's datasets.

Features Importance Analysis. To address RQ2, we analyzed the feature importance scores derived from the optimal Random Forest configurations. This model type inherently provides a widely used metric for feature importance, known as Gini importance. This metric quantifies the average decrease in node impurity achieved by splitting on a specific feature, reflecting the extent to which each feature contributes to the model's predictive power. For each political candidate, we aggregated the Gini scores obtained in the five optimal Random Forest configurations identified during model training (Sect. 3.4) to obtain a more robust result.

4 Results

4.1 Data Collection

The dataset encompasses various features as outlined in study [8], with most being categorical except for the continuous *Duration* feature. These features include the *Candidate* (Lula or Bolsonaro), *Duration* (length) of the video, *Aristotelian Rhetoric* (logos, ethos, pathos), *Content Type* (such as personal, political ideological, campaign act, and trend), *Functional Approach* (attack, defense, acclaim), *Tonality* (positive, negative, neutral), and *Main Character* (candidate, candidate with voters, voters, influencers, opponent). These features were manually labeled based on a predefined taxonomy aimed at classifying varied aspects of the politicians' discourse displayed in each video. For example, *Aristotelian Rhetoric* consists of three categories (values) of rhetorical appeal, which aim to convince the interlocutor: *logos*, which indicates an appeal to logical reasoning and argumentative structure; *ethos*, appeal to the author's character, such as their trustworthiness and qualifications; and *pathos*, appeal to the audience's emotions. For further details on the attributes, please refer to study [8].

Additional features used in the present study, which were collected as part of the research study [8] but not reported in it, are described below:

- **Rhetorical Device** - with the following categories: *advertisement* – videos (TikTok posts) that aim to promote or sell the political candidacy; *call to action* – seek to mobilize the public to take a specific action, such as voting or participating in an event; *collective appeal* – appeal to the union or collaboration of a specific group or society in general; *commitment* – display the politician at an event such as parades and speeches; *endorsement* – represent support for the politician and a way to value

their ideas and what they represent; *fact/statistic* – present facts or statistics to inform or persuade the public; *humor* – use humor to convey a political message; *opinion* – the politician expresses their opinion; *personal appeal* – the politician appeals to empathy or personal identification; *political statement* – the politician makes a clear statement on a political issue; *thanks* –the politician expresses gratitude to specific people or groups; *urgency* – videos that convey a message of urgency regarding a cause or political issue; *none* – videos that do not fit into any of the listed categories.

- **Textual Content** - has four categories that represent how the video is accompanied by a caption: *hashtag* (only), *none* (no text), *text*, *text* + *hashtag*.
- **Post Date**: the date in which the video was posted on TikTok. The data collection period spanned from June 30th to October 30th, 2022.
- **Collection Date**: the date on which all the post's features were collected. Approximately 75% of the posts were collected in October the 2nd, while the remaining data was collected on October the 30th, 2022.

Additionally, we had access to engagement variables for each post (e.g., Plays, Shares, Comments), but we focused on *Likes* as explained in Sect. 3.2.

4.2 Preprocessing

During data cleaning, eight posts were removed due to missing values, all from Bolsonaro's posts. The final data was then separated into two datasets, one for each candidate, resulting in 261 posts in Bolsonaro's dataset, and 308 posts in Lula's dataset.

The target attribute for binary classification was created by performing a median split on the number of likes. This resulted in two categories: *high* engagement posts and *low* engagement posts. In Bolsonaro's dataset, the median number of likes was 31,600. Consequently, 131 posts were classified as low engagement and 130 posts as high engagement. For Lula's posts, the median was 19,400, resulting in 154 posts categorized as *low* engagement and 154 posts categorized as *high* engagement.

We engineered a new feature called *Days Elapsed* by subtracting the publication date of each post from the date of data collection. This feature accounts for the varying exposure times of posts in our dataset and for the different periods of campaign that it was related to, which could potentially influence engagement levels.

4.3 Model Training

The models had their parameters selected and their performance assessed using the nested 5-fold cross-validation procedure explained in the Sect. 3.3. For each candidate, considering all options for all stages of the pipeline, including possible data transformations and hyperparameters, a total of 243 different configurations of the MLP model were tested, along with 360 configurations of SVM, 216 of Random Forest, 216 of Logistic Regression, and 144 of KNN. These numbers reflect the various combinations of feature transformations and hyperparameter settings explored for each model type. Considering the two levels of 5-fold cross-validation (25 iterations in total) and the final training with the best configuration for each fold (5 additional trainings, being one per outer fold), this resulted in 29,480 model training instances being performed. The results of these training runs are discussed in detail in the following sections.

4.4 Data Transformation

This section analyzes the data transformations chosen in the best-performing model configurations found in each of the five outer folds of nested cross-validation employed. The data transformations are presented in Table 2, where the first three transformations (rows of the upper part of the table) are categorical, while the last three transformations (lower part) are numerical. Note that there is a total of five categorical transformations and five numerical transformations per model (column). Our primary focus is on identifying the most frequently selected transformations for categorical data, as these constitute a significant portion of the features in our datasets.

Table 2. Data transformations chosen in the model's best configurations.

	Bolsonaro						Lula					
	MLP	SVM	RF	LR	KNN	*Total*	MLP	SVM	RF	LR	KNN	*Total*
one-hot	1	2	1	1	4	*9*	0	4	0	3	2	*9*
target/classif	2	2	2	1	1	*8*	1	0	1	1	1	*4*
target/regr	2	1	2	3	0	*8*	4	1	4	1	2	*12*
min-max	2	2	NA	2	4	*10*	5	1	NA	3	2	*11*
standard	3	3	NA	3	0	*9*	0	0	NA	2	2	*4*
identify	0	0	NA	0	1	*1*	0	4	NA	0	1	*5*

Categorical Features. For Bolsonaro's posts, the optimal configurations across all model types exhibited a relatively even distribution of chosen transformations. One-hot encoding was selected in 9 configurations, proving particularly useful for KNN models (being chosen in 4 of the 5 best configuration). Both target encoding for binary classification and target encoding for regression were chosen 8 times each. Similarly, in Lula's dataset, one-hot encoding was favored in 9 of the best configurations, especially for Logistic Regression models (3 out of 5 configurations). Target encoding for binary classification was chosen 4 times. Target encoding for regression was selected 12 times, being present in at least one optimal configuration of each model type.

Overall, target encoding for regression emerged as the most frequently chosen transformation for categorical data (20 times, being 8 in Bolsonaro's dataset and 12 in Lula's dataset). This supports our hypothesis that encoding categorical features based on the distribution of the raw continuous variable (number of likes) can capture valuable information even when the target variable is ultimately transformed in a 2-class variable (*high* or *low* engagement).

Numerical Features. Min-max scaling was the most common transformation for numerical features, being found 21 times in the optimal model configurations. Standard scaling was also frequently selected, 13 times. No transformation (*identity*) was chosen in 6 model configurations, most of them in models trained on Lula's data.

4.5 Analysis

Model Performance – Bolsonaro's Posts. The performance results for Bolsonaro's dataset are presented in Table 3. Support Vector Machine (SVM) emerged as the best-performing model with an F1-score of 0.69. Logistic Regression, K-Nearest Neighbors (KNN), and Multi-Layer Perceptron (MLP) followed closely with F1-scores of 0.68, 0.67, and 0.66, respectively. Random Forest achieved a slightly lower F1-score of 0.64.

Table 3. Performance results for the Bolsonaro dataset (mean ± standard deviation).

Model	Accuracy	Precision	Recall	F1-score
SVM	**0.693 ± 0.043**	0.688 ± 0.038	**0.707 ± 0.110**	**0.693 ± 0.060**
LR	0.685 ± 0.024	0.686 ± 0.038	0.692 ± 0.087	0.684 ± 0.039
KNN	0.678 ± 0.053	**0.695 ± 0.069**	0.653 ± 0.048	0.670 ± 0.036
MLP	0.659 ± 0.040	0.655 ± 0.043	0.669 ± 0.071	0.660 ± 0.046
RF	0.655 ± 0.052	0.672 ± 0.078	0.661 ± 0.163	0.647 ± 0.073

While the overall model performance may not be considered remarkable, the results demonstrate the potential for using machine learning to identify posts with a high probability of achieving high engagement. For example, the SVM model achieved a recall of 0,707, indicating that it can correctly identify actual high-engagement posts 41% better than a random classifier (that would display a recall of 0.5). The precision of 0.688 indicates that when it positively predicts that a post will have high engagement, the model is correct 68.8% of the time, which is a 37.6% improvement over a random classifier (precision 0.5).

Model Performance – Lula's Posts. The performance results for Lula's dataset, presented in Table 4, indicate that the models struggled to achieve satisfactory performance. This suggests that the selected features may not be sufficient to effectively differentiate between high and low engagement posts in Lula's case. The SVM model achieved the highest F1-score of 0.61 (with a standard deviation of 0.07), primarily due to a high recall of 0.85 but a low precision of 0.50. This suggests that the model can identify a large portion of actual high-engagement posts but also generates a substantial number of false positives. Furthermore, three of the five model types (Logistic Regression, Random Forest and MLP) achieved F1-scores below 0.50, indicating their limited ability to effectively distinguish between the two classes in Lula's case.

Features Importances. We also analyzed, for each candidate, the feature importances identified in the 5 best Random Forest models (one per fold of the outer cross-validation). The exact names of the features may vary between models, depending on the transformations applied to the categorical features. One-hot encoding transforms one feature into multiple features, one for each value, that appear named as "<feature>_<value>" in the results. However, target encoding preserves the name of the features. For the models that used the same transformations and used the same feature names, we averaged the importances of these models.

Table 4. Performance results for the Lula dataset (mean ± standard deviation).

Model	Accuracy	Precision	Recall	F1-score
SVM	0.499 ± 0.024	0.504 ± 0.020	**0.858 ± 0.222**	**0.618 ± 0.074**
KNN	**0.548 ± 0.030**	**0.552 ± 0.027**	0.519 ± 0.043	0.534 ± 0.030
LR	0.500 ± 0.023	0.398 ± 0.200	0.612 ± 0.382	0.469 ± 0.250
RF	0.483 ± 0.070	0.493 ± 0.070	0.448 ± 0.110	0.459 ± 0.075
MLP	0.509 ± 0.015	0.405 ± 0.202	0.625 ± 0.436	0.457 ± 0.274

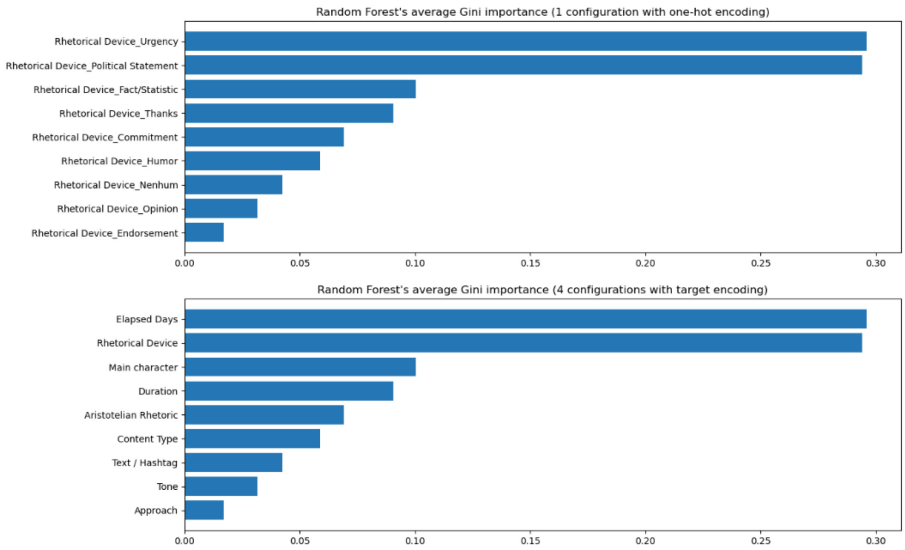

Fig. 1. Importance of the Bolsonaro candidate's features.

From the 5 best configurations of Random Forest classifier trained in Bolsonaro's data, only one of them applied one-hot encoding. The feature importances calculated from this model for the best features is reported in the upper part of Fig. 1. This model highlights the significance of specific Rhetorical Devices in driving engagement, with urgency and political statement emerging as particularly influential categories.

The remaining four models adopted target encoding and the average value of the feature importances extracted from these models is shown on the bottom part of the same figure. While these models also identified *Rhetorical Device* as a crucial feature, they did not provide specific insights into the impact of individual categories. Interestingly, *Elapsed Days*, a feature unrelated to speech content, emerged as the most important feature, indicating the significant role of post timing within the campaign period. The other features present notably lower scores.

All the five best configurations of the Random Forest model trained with Lula's data employed some form of target encoding. The averaged importances extracted from these models are shown in Fig. 2, ordered from highest to lowest non-zero importance. When compared to Bolsonaro's results for target encoding configurations, we observe a remarkable similarity between the rankings of the features. The only difference in the order is the fifth and sixth positions, which appear swapped. This similarity in the feature importance's rankings demonstrates consistency in the relevance of these features for predicting high-engagement posts for any candidate.

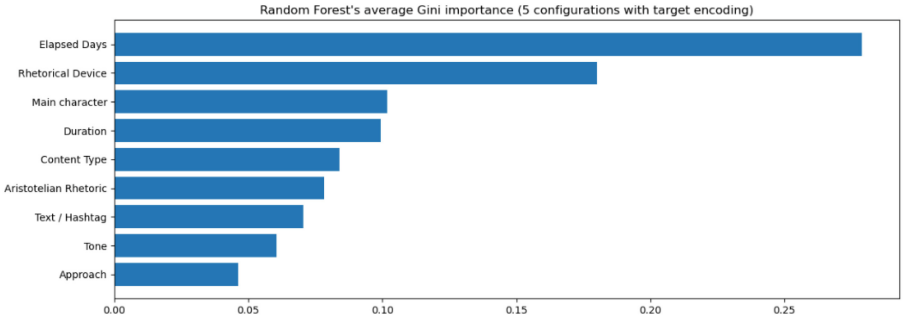

Fig. 2. Importance of the Lula candidate's features.

5 Discussion

What makes a post engaging can be a series of things, not just the type and content of the post, including everything from the general audience's taste to whether it became a 'meme' or got involved in controversy in some way. Thus, given the limited aspects captured by the data utilized, we believe our results were relevant. The results for Bolsonaro demonstrate the potential for using machine learning to identify posts with a high probability of achieving high engagement with an F1-score of 0.70. The SVM model can correctly identify actual high-engagement posts 41% better than a random classifier (0.707/0.5), and the precision metric indicates that when it positively predicts that a post will have high engagement, it has a 37.6% higher success rate than a random classifier (0.688/0.5). In a typical practical application for this model, such as assisting a candidate's social media team in filtering content to be posted, the model's ability to identify potential high-engagement posts could be valuable, potentially improving expected engagement compared to not using such a filter. It is important to note that false positives or false negatives would not be particularly harmful in this application. Therefore, this performance is sufficient and represents an improvement over the absence of machine learning-based filtering.

The models trained on Lula's data exhibited comparatively lower performance with an F1-score of 0.62, reducing their potential for application. This discrepancy indicates that the factors influencing engagement for Lula's audience are more complex. Further

research is necessary to explore these complexities, identify new features and refine the models accordingly.

The feature importance analysis identified two features with dominant roles: Rhetorical Devices and Elapsed Days. The inclusion of the Elapsed Days feature was crucial for controlling variability in the data, as posts had different durations to accumulate likes. However, we acknowledge that in a practical application aimed at predicting the engagement of a new post, this feature would not be available. To address this, a potential solution could involve allowing users to input a desired timeframe for the prediction (e.g., 'how many likes to expect in X days'). This would enable the model to adjust its predictions based on the specified time horizon. Additionally, future research could explore training models that do not rely on Elapsed days or that incorporate user-defined time projections, making them more applicable to real-world scenarios. In general, dimensionality reduction was not explored in this work and may also be investigated in future works.

Overall, we believe that this analysis is a starting point for understanding the factors driving engagement on his social media content. Further investigation into the specific rhetorical devices, content type preferences, and feature interactions can offer deeper insights and inform effective communication strategies. We could also employ other methods to extract information about feature importances.

For a critical assessment of this study, it's crucial to note potential methodological biases that may influence our result or its interpretation. Manual post categorization, despite rigorous use of multiple classifiers, may be subject to subjectivity in applying criteria. Moreover, data collected during the 2022 Brazilian election campaign introduces contextual variations among candidates, limiting findings' generalizability to other electoral contexts or periods. Thus, while valuable for practical use, their applicability beyond the specific context warrants caution.

6 Conclusion and Future Works

This article presents a study aimed at predicting the engagement of TikTok posts made by the two main candidates in the 2022 Brazilian elections: Lula, the current president of Brazil, and Bolsonaro. The method involved acquiring the dataset from a previous work [8], followed by data analysis, preprocessing, and training. The study accessed additional variables not utilized in the work and engineered a new feature, enabling novel analyses and approaches. The dataset was cleaned and transformed, with special attention given to the categorical features, which were abundant. The transformed data was used to train and compare five types of classification models optimized by a grid search with nested cross-validation: Support Vector Machine (SVM), Logistic Regression, KNN, MLP Neural Network, and Random Forest. These models were evaluated based on their performance metrics, followed by an analysis of feature importance for each candidate.

The classification problem and dataset presented various challenges. For Bolsonaro's posts, the SVM model performed best, achieving an F1-score of 69%, with a recall of 70% and precision of 68.8%. These results suggest that machine learning can effectively identify posts likely to gain high engagement, which could be valuable for social media teams in decision-making. This is because accurate predictions can lead to positive

outcomes, while the impact of incorrect predictions in this context is relatively low. However, for Lula's posts, models struggled, with the highest F1-score of 61% from the SVM model, which had a high recall (85%) but low precision (50%), indicating a significant number of false positives. The most important features for classifying the posts were also identified for each candidate, validating the relevance of the discourse features in effectively predicting engagement.

We also discussed limitations that highlight the need for further investigation, such as investigating the removal of the 'Elapsed Days' feature and exploring the effect of reducing the number of input features used by the models. Future work could also involve using data from other social media platforms, exploring computer vision techniques (applied to post images), text mining, and additional machine learning methods. Lastly, since content analysis was performed manually to create the dataset from [8], employing computer vision techniques, large language models (LLMs), or automatic post classification offers promising directions for future advancements.

The findings contribute to understanding the evolving dynamics of publications on the TikTok platform within the context of political profiles, especially as the social media continues to grow in Brazil. Additionally, the results hold relevance beyond social network analysis, impacting areas such as social and political sciences by providing insights into politicians' posting dynamics and voters' behavior and interests on this platform. Furthermore, the study offers valuable information for the fields of marketing and advertising, aiding decision-making regarding publication content based on the political profile of the author, their electorate, and the target audience. The results also empower citizens by raising awareness of the content displayed on their screens, preventing political manipulation, and fostering a more autonomous, secure, and healthy environment for political debate.

References

1. DataReportal: DIGITAL 2024: BRAZIL. https://datareportal.com/reports/digital-2024-brazil. Accessed 24 June 2024
2. Brito, K., Paula, N., Fernandes, M., Meira, S.: Social media and presidential campaigns – preliminary results of the 2018 Brazilian presidential election. In: Proceedings of the 20th Annual International Conference on Digital Government Research, pp. 332–341. ACM, New York (2019). https://doi.org/10.1145/3325112.3325252
3. Brito, K., de Lemos Meira, S.R., Adeodato, P.J.L.: Correlations of social media performance and electoral results in Brazilian presidential elections. Inf. Polity **26**, 417–439 (2021). https://doi.org/10.3233/IP-210315
4. Brito, K., Filho, R.L.C.S., Adeodato, P.J.L.: A systematic review of predicting elections based on social media data: research challenges and future directions. IEEE Trans. Comput. Soc. Syst. **8**, 819–843 (2021). https://doi.org/10.1109/TCSS.2021.3063660
5. Heiss, R., Schmuck, D., Matthes, J.: What drives interaction in political actors' Facebook posts? Profile and content predictors of user engagement and political actors' reactions. Inf. Commun. Soc. **22**, 1497–1513 (2019). https://doi.org/10.1080/1369118X.2018.1445273
6. Xu, L., Yan, X., Zhang, Z.: Research on the causes of the "Tik Tok" app becoming popular and the existing problems. J. Adv. Manag. Sci. 59–63 (2019)
7. Simon Kemp: Digital 2022: Brazil. https://datareportal.com/reports/digital-2022-brazil. Accessed 03 June 2024

8. Santana, M., Lima, J., Correa, A., Brito, K.: Engajamento no TikTok dos candidatos às eleições Brasileiras de 2022 – Resultados Iniciais. In: Anais do XII Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2023), pp. 151–162. Sociedade Brasileira de Computação - SBC (2023). https://doi.org/10.5753/brasnam.2023.230641
9. Bimber, B.: Digital media in the Obama campaigns of 2008 and 2012: adaptation to the personalized political communication environment. J. Inform. Tech. Polit. **11**, 130–150 (2014). https://doi.org/10.1080/19331681.2014.895691
10. Francia, P.L.: Free media and twitter in the 2016 presidential election. Soc. Sci. Comput. Rev. **36**, 440–455 (2018). https://doi.org/10.1177/0894439317730302
11. Tumasjan, A., Sprenger, T., Sandner, P., Welpe, I.: Predicting elections with twitter: what 140 characters reveal about political sentiment. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 4, pp. 178–185 (2010). https://doi.org/10.1609/icwsm.v4i1.14009.
12. Brito, K., Silva Filho, R.L.C., Adeodato, P.J.L.: Stop trying to predict elections only with twitter – there are other data sources and technical issues to be improved. Gov. Inf. Q. **41**, 101899 (2024). https://doi.org/10.1016/j.giq.2023.101899
13. Oueslati, O., Hajhmida, M. Ben, Ounelli, H., Cambria, E.: Sentiment Analysis of Influential Messages for Political Election Forecasting. Presented at the (2023)
14. Vepsäläinen, T., Li, H., Suomi, R.: The role of social media platforms in forecasting elections: a comparison of twitter and facebook. Digit. Gov. Res. Pract. (2024). https://doi.org/10.1145/3651227
15. Brito, K., Silva Filho, R.L.C., Adeodato, P.: Please stop trying to predict elections only with Twitter. In: DG.O 2022: The 23rd Annual International Conference on Digital Government Research, pp. 88–95. ACM, New York (2022). https://doi.org/10.1145/3543434.3543648
16. Simon Kemp: Digital 2022: Global Overview Report. https://datareportal.com/reports/digital-2022-global-overview-report. Accessed 03 June 2024
17. Lima, J., Santana, M., Correa, A., Brito, K.: Dataset for The use and impact of TikTok in the 2022 Brazilian presidential election (2023). https://doi.org/10.7910/DVN/9L7LEI
18. Lima, J., Santana, M., Correa, A., Brito, K.: The use and impact of TikTok in the 2022 Brazilian presidential election. In: Proceedings of the 24th Annual International Conference on Digital Government Research, pp. 144–152. ACM, New York (2023). https://doi.org/10.1145/3598469.3598485
19. Pargent, F., Pfisterer, F., Thomas, J., Bischl, B.: Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. Comput. Stat. **37**, 2671–2692 (2022). https://doi.org/10.1007/s00180-022-01207-6