The digital behavior of voters in interactions with the social media posts of candidates running for elections

Author: Heriberto Alexandre da Silva Filho

The extensive use of digital tools and digital marketing strategies over the last few years has become increasingly more frequent and characteristic in political campaigns Within this scenario, this study aims to investigate the use of SM in contemporary political communication, seeking to understand the features that influence the engagement of voters in posts by politicians on their social media profiles. As a case study, we have focused on the Brazilian presidential election in 2018. The investigation was based on an analysis of politicians' posts on Instagram, Twitter, and Facebook (N = 1319) in the last two weeks before the elections, which investigated features such as functional approach, the Aristotelian rhetoric adopted, and the type of content, among others, and established relationships between these features and user engagement. This study also proposes to investigate the feasibility of using machine learning models to predict the level of engagement of the candidate's posts. Finally, another objective of this paper is to find similarities or differences between the digital campaign strategies, and their impacts on the level of engagement, of the two candidates with the best electoral results. Our main results indicate that the platform with the highest level of engagement was Instagram, together with polarized discourses that presented speeches of attack and defense or emotionally charged topics tended to engage more. Regarding the predictions, the Gradient Boosting model proved to be efficient, R^2 =0.77, to make the predictions. Regarding the digital campaign strategies, although the two candidates are from opposite political sides, it was possible to find more similarities, such as: functional approach, content structure, and content type, and others...than differences. However the few differences found also represent a valuable result for the understanding of the political landscape, there were divergences for example in Aristotelian rhetoric, content type, and rhetorical device. All these results helped to understand how the electorate interacts with the candidates' speeches in a new era of digital campaigning.

Additional Keywords and Phrases: Data Analysis, Machine Learning, Artificial Intelligence, Social Networks, social media, Elections, Brazil, Facebook, Twitter, Instagram

1. INTRODUCTION

Digital tools, especially social media, are playing an important role in political campaigns worldwide, since they are used as mediators by politicians, who aim to spread their messages and public policies, and by the electorate, who seek to become more informed and aware of their representatives [1]. This creates a more democratic scenario more prone to social-political dialogue and promoting political participation [2]. However, the phenomenon of social media and their use in political campaigns is a very recent development, when compared to the history of social-political construction, since the most used social media, such as Facebook and Twitter, were only launched to the public in 2006, and Instagram emerged in 2010.

It may also be affirmed that the insertion of social media into the political scenario has brought a series of new features to the political panorama, since their use has provided a direct, instantaneous communication channel, thereby breaking with logistical, geographical, and even financial restrictions. Another important point regarding the use of social media is their bivalency concerning their power of coverage, since they may be used as a channel of mass communication, seeking to reach new audiences, but may also be used to strengthen and consolidate the relationship between politicians and those who already follow them, thereby promoting a more direct, intimate relationship. This direct engagement can influence political decision-making since political actors can derive the climate of opinion from user engagement and are confronted with direct feedback from potential voters and opinion leaders [3].

Once social media became included in electoral campaigns and the political scene, thereby helping to break the previously mentioned barriers and restrictions, it may be affirmed that the political scenario, especially about elections, was strongly impacted. It is possible to highlight two of these major impacts. The first is the democratization of gaining access to information. In the 2016 U.S. presidential election, SM was the most important source for young people to obtain campaign information, and an increasing number of Facebook users have been retrieving their political information directly from the candidates' Facebook pages [4]. Another relevant impact has been the equalization of electoral political participation since social media offers attractive means, is cost-beneficial, and leads to the emergence of new representative figures who previously had no visibility via the traditional media. One clear example of this was the 2018 Brazilian presidential election when the candidate with the highest level of social media engagement, although with little exposure to traditional media, was elected [5]. Such impacts, among others, may affect the political behavior of the electorate [6]– [9].

In this context, much academic research has been devoted to directly correlating the performance of politicians on social media and their electoral performance. Thus, researchers have attempted to build models in order to predict electoral results based on social media data [10][11]. Moreover, in Brazil, a direct correlation between social media performances and electoral performance has already been identified [12]. However, this research area is still in its initial stages, and yet, there is no common consensus in the literature regarding what drives the interaction of votes on the social media posts of candidates.

In this paper, we explore the factors that help to explain what distinguishes a candidate's post with a high level of engagement from one without engagement, to advance our understanding of the dynamics created by an increasingly interactive online campaign environment. Some of these factors are related to the content of the post, i.e., the political theme/issue under discussion, for example, "Can a distinction be made between posts on political-propositional issues and those on political-ideological issues?". There are also factors related to how the theme is transmitted, for example: "Do posts with more offensive speeches and attacks engage more than passive posts?". Lastly, we also investigate whether any digital factors may impact the engagement of the post, such as the day and time of publication, the social media used, or the type of post (image, hyperlink, text...). Thus, by answering these questions, we may obtain the complete context into which the message has been inserted. With the knowledge of which factors do or do not impact the engagement of a post, we also set out to use machine learning models to predict the level of engagement of a post given the presence or absence of these factors. And finally, we set out to find differences and/or similarities between the digital strategies of the two candidates with the best election results and opposing political spectrums.

As a case study, we obtained 1,319 posts from candidates in the 2018 presidential election in Brazil. We then performed an analysis where each post was classified into different features according to our proposed taxonomy, which was designed from a literature review of the taxonomies of the main studies that have also addressed the theme of engagement in posts by politicians on SM. Once the posts were classified, we calculated the descriptive data on the categories of posts, such as count, mean, max, min, and standard deviation, and performed static tests (t-unpaired test) to identify the features that influence the engagement of the posts. And finally, we submitted this dataset to different machine learning models considering 3 different scenarios: 1- data referring to the posts from all the 13 candidates, 2- only the data referring to the posts from candidate Jair Bolsonaro, first place in the electoral results, and lastly 3- only the data referring to the posts from candidate Fernando Haddad, second place in the electoral results.

The results of this study are a contribution to research in this area in several manners. First, to the best of our knowledge, this is the first study to analyze and correlate the results of many different social media platforms, including Facebook, Twitter, and Instagram. We have also proposed and applied a new taxonomy, based on the most relevant previously defined features. Lastly, the results are presented by applying our methodology to a recent election in a country with a different electoral scenario to that of the U.S., which is generally the most studied country.

The remainder of this paper is organized as follows: Section 2 presents the background to this research, including general background, a discussion on the main related work, and a brief overview of the 2018 Brazilian presidential elections. In Section 3, we detail the methodology adopted for this research. Section 4 presents and discusses the results, followed by Section 5, which presents the threats to validity, and finally with have Section 6 which contains the conclusions and final considerations of the research.

2. BACKGROUND

Currently, there is a strong presence of SM in the political scene since candidates have begun to adopt social media as the main item of their campaign strategies. The prime example of a presidential campaign conducted explicitly and substantially using social media was the 2008 Barack Obama campaign in the U.S [13], [14]. With the success of the campaign that led to the victory of the first black president of the U.S., the use of social media became a model, and thus influenced subsequent elections, such as that of Trump in the 2016 U.S. presidential elections [16] and Brexit [15]. In the 2018 Brazilian presidential election, Bolsonaro had almost no TV time, but his SM profiles received more than half (55%) of the SM interactions received by all the other candidates, and he was elected [5]. On the other hand, the candidate with the largest amount of TV time and fewer SM followers ended up in fourth place.

The influence of social media within the political scene, especially in elections, as highlighted above, has been possible because of the results they brought both for the politicians and their followers. Several studies have demonstrated that engagement in activities such as "like" and commenting on candidate social media posts are significantly related to political engagement among supporters [17]– [19]. Considering that activities such as commenting on a candidate's Facebook pages also predict positive attitudes toward candidates [20], this relationship suggests that user response represents a tangible style of potential mobilization for candidates, thereby further identifying and consolidating supporter engagement. Indeed, recent studies have presented high correlations between the performance of candidates on SM and their electoral performance [12], and the possibility of predicting electoral results based on SM data [10], [11]

2.1 RELATED Works

In 2013, when the mass popularization of contemporaneous SM began, [21] attempted to understand the relationship between the presence of emotional charge contained in posts made on social media and

information sharing between users. For this, they used a dataset with more than 165,000 posts from Twitter and examined whether the sentiment occurring in the social media content was associated with the information-sharing behavior of a user. With this, it was possible to cross-check the classification of the post with its number of retweets (Twitter's native sharing functionality) and understand whether there was an influential relationship between two pieces of information: the emotional charge and sharing. The study concluded that emotionally charged posts are shared increasingly more quickly than neutral posts.

In 2017, [22] proposed a systematic analysis through an investigation of user reactions to posts by political candidates via Facebook. Some of the targets of the research were: to discover the extent to which contextual factors, such as race, candidate, district, and state characteristics, influenced user engagement in candidate posts on SM, and how variations in the content of the candidate posts on SM corresponded to variations in user responses. This investigation revealed notable patterns regarding candidate and supporter interactions on SM, and that some variations of these posts, such as timing and content, were significantly related to how users engaged with the post through likes and comments. Although they took into consideration important features regarding the post, such as tone and campaign context, the total set of variations chosen for the analysis was somewhat reduced and did not reflect the complexity related to the problematics of the theme. The research was also restricted to Facebook posts only.

Bene's study [23] aimed to investigate how communication occurs between political actors and their stakeholders in the political scenario. The research focused on the historical and temporal cross-section of Barack Obama's campaign on Facebook during the period leading up to the 2012 presidential elections. From this context, it was not only possible to obtain the posts, but also the reactions and interactions of users to the posts through Facebook's native functionality (like, comment, and share). It was then possible to cross-reference this information, posts, and interactions, to conduct the analyses, which among other conclusions, reported that: (i) followers selectively engaged with the campaign messages and often interacted more with policy-oriented messages than with promotional messages; and (ii) the Obama campaign used Facebook as a top-down promotional tool, focusing on Obama's personality and as a means of strategically guiding followers to act. In terms of limitations, we would mention the very restricted context that was analyzed, since it only included a single political actor, Obama, and a single social media, Facebook.

Heiss [3] investigated the features, on the level of profile and post, that lead to user engagement on social media posts of political actors. The engagement was considered as the number of likes, shares, and comments. A database of Facebook posts by political actors (N = 1915) was used to undertake the analyses. The results suggest that the post level, reasoning, post length, and quotes to competing political actors have the potential of increasing different types of user engagement. Moreover, negative intonation increased user engagement, and positive emotional expressions had a stronger impact on user engagement than negative emotions. In addition, humorous messages were more likely to be commented on, liked, or shared, while mobilizing suggestions brought about predominantly negative effects on user engagement. The main innovation of this paper is the investigation into the reasons why political figures react to user comments on their posts. The main limitation of the study was common to most of the previous works, the focus was on just one social media platform.

Lastly, and considering the Brazilian elections, as background to their research, [5] used the presidential election of Brazil in 2018 to analyze and investigate the relationship between the electoral result of the politician and his or her social media presence/performance. To investigate this relationship, the candidate's performance on social media and the votes received the numbers of followers, and all posts from all 13 presidential candidates on the three major social medias, Facebook, Twitter, and Instagram, from January-October 2018 were collected and analyzed. Some of the results presented in the paper are: (i) Politics usually focused on engaging words and avoided sensitive topics; (ii) Posts made on Instagram tended to have stronger engagement; and (iii) There were some strong correlations between votes and followers, and votes and engagement, mainly on Instagram. The strengths of this research include the use of an extensive database containing posts from different social media platforms over a long period of time, 10 months before the election. However, they were unable to identify which post characteristics were most correlated with the interactions they received.

It may be surmised that social media already plays an important role in the political arena and that it is being used as an engagement tool between politicians and their followers. It may also be concluded that there are already relevant studies that seek to identify which factors are responsible for engaging a post. However, we may highlight some points of improvement which are in common among the studies, and which may be improved by this or further research. The main point is that, except for [5], all studies were restricted to a single SM platform. As a result, large sets of data and indicators of other platforms are ignored. For example, the same person often publishes the same content on many platforms and receives different levels of engagement on each. This may be due to the difference in the users of platforms: the same content may be more appealing for users on one platform than for those on another. Thus, we have opted to broaden our

research and contemplate more than one social media and thus extend the knowledge on user behavior on social media in general.

2.2 A Brief overview of the 2018 Brazilian elections

The 2018 Brazilian presidential election featured 13 candidates and was marked by a fervorous contest between two of them, Jair Bolsonaro and Fernando Haddad. The following is a brief overview of all the candidates.

Jair Bolsonaro, elected president in this election, already had a long political career behind him, since he had been a Federal Deputy since 1991. Bolsonaro emerged mainly due to the movement against the Workers' Party (PT), the party that had previously presided over the country for the last 13 years, with its last representative Dilma Rousseff being impeached. Well-known for his politically incorrect opinions, he opposed the ideas of the Workers Party and its two main representatives Luiz Inácio Lula da Silva and Fernando Haddad and presented a right-wing economic proposal. During the campaign, he was stabbed in the stomach while interacting with supporters. His condition prevented him from returning to public activities and debates for the remainder of the first round. To some extent, this favored the adoption of social media and digital media, such as the live broadcasts during his campaign, even from the hospital.

Fernando Haddad was the former mayor of São Paulo and was the Workers' Party (PT) candidate. His party had won the previous four presidential elections, but certain party members and government entities had been involved in many corruption scandals. Economically, the party is left-oriented and strongly aligned with human and minority rights. Haddad took over the candidacy after Lula, the most likely candidate to run in the elections for the PT, and indicated by many polls as being the favorite, had his candidacy rejected by the Superior Electoral Court due to corruption allegations.

Ciro Gomes (center-left) and Geraldo Alckmin (center-right) presented themselves as moderate options for left and right-wing voters. João Amoêdo, a right-wing businessman, was the "non-political candidate." Cabo Daciolo (far-right) was often the "comic candidate." Henrique Meirelles (center-right) represented the current government of the time, which was very unpopular due to the impeachment of the last president. Marina Silva (center-left), the third most favored candidate in 2014, and Alvaro Dias (center-right) completed the list of the "third-way" candidates. Guilherme Boulos (ultra-left), Vera Lúcia (ultra-left), Eymael (center-right), and João Goulart Filho (center-left) composed a group of candidates with just a few supporters [5].

3. METHODOLOGY

Considering the context presented in Section 2, this paper aims to analyze the posts made by election candidates on social media and answer the following research questions:

RQ1. Which features of social media posts have the greatest influence over user engagement on the social media profiles of candidates in the run-up to elections?

RQ2. Is it possible to use Machine Learning models to predict the engagement of politicians' social media posts in the run-up to elections?

RQ3. Considering the Brazilian political scenario in the 2018 elections, it is possible to identify differences in digital strategies, between the two candidates with the best electoral results and from opposite ideological political spectrums?

The methodology employed to answer the RQ1 consists of three steps: (i) to revise and define a taxonomy for the classification of posts; (ii) to collect related data from SM platforms and classify them according to the defined methodology; and (iii) to analyze collected data in two ways, through descriptive analysis and by performing statistical tests. Another proposal presented in this paper consists of applying machine learning models to predict the level of engagement of a post made by political candidates on their social networks during the election period, this proposal is represented on RQ2. The methodology employed to answer the RQ2 consists of the following steps: (i) Define the machine learning models to be used in the predictions (ii) Define the parameterization and settings for each of the machine learning models selected in the previous step. (ii) Train the machine learning models, configured according to the previous step, using the posts collected from the social networks and their classifications according to the taxonomy defined in this paper. Our last research question, RQ3, tries to identify similarities or differences in digital strategies about the level of engagement on social networks considering the scenario defined in this research to answer the RQ3 we separated the data from the two candidates, Jair Bolsonaro e Fernando Haddad and trained the machine learning models for each of them separately.

As a case study, we have focused on the 2018 Brazilian presidential election. Thus, we collected and analyzed data from all 13 candidates running for the first round of the election. We considered the three main SM platforms in Brazil, Facebook, Twitter, and Instagram, and the last two weeks before the first round of the elections. Details of each step of the methodology are presented below.

3.1 Data Collection

The dataset used for this study is the same as that used by [12], which they kindly provided. The dataset contains a total of 44,265 posts made by the 13 candidates on Facebook (12,776 posts, 29%), Twitter (23,312 posts, 53%), and Instagram (8,177, 18%), from January 1, 2018, to October 6, 2018 – one day before the first round of the election. These posts generated 290 million interactions –143 million on Facebook (49%), 32 million on Twitter (11%), and 116 million on Instagram (40%). For this study, we considered the two weeks prior to election day since this is the period with the highest candidate and follower activity and is often considered decisive for the elections. Thus, we analyzed a subset of this dataset, containing 1,319 posts with interactions of voters.

To represent the level of engagement of a post we chose to use the metric of the number of likes on that post. This choice was based on 3 arguments. The first one, when compared to the other metrics of sharing and commenting, we observed that the number of likes on a post has more potential to engage the post than the number of comments or the number of shares. This conclusion was reached by performing a correlation test using Pearson's technique. Pearson's correlation coefficient (r) is the most common way to measure a linear correlation. It is a number between -1 and 1 that measures the strength and direction of the relationship between two variables. In this case, we compared the impact of the three various with the overall engagement level of the post and obtained the following results:

Table 1: Pearson Correlations between the 3 metrics and engagement

Pearson Correlations			
1	+0.939	Likes	Engagement
2	+0.711	Comments	Engagement
3	+0.672	Shares	Engagement

Second, for each user on the social network, it is only possible to like the post once, unlike the other metrics obtained by the functionalities: comment and share, which can be replicated n times for each user, that is, a single user can comment/share the same post 1,2,3... "n" times. With this, we can consider that the metric of the number of likes more truly reflects the impact of the engagement of the post on the network of users on social medias. Third, the fact that the 3 social networks analyzed share this same functionality, like a post, which is not the case with the other functionalities such as sharing and commenting. On Instagram, it is not possible to share a publication, just as it is not possible to comment on a Twitter post, since on Twitter the comment generates a kind of posting sharing, called by the platform as "retweet".

3.2 Taxonomy for post-classification

The first step before trying to identify which features have higher correlations with engagement on candidate posts is to define the features. First, we reviewed the literature related to classifying posts on SM [3], [21]–[26], not necessarily in the political context, in which we analyzed the set of previously proposed taxonomies. Next, we applied the following process:

First, to list all taxonomies, placing all the features of the five studies together and forming an extremely extensive set of classifications. This initial step was performed to allow us to analyze all possible classifications.

Second, to eliminate features outside the electoral context and unrelated to our research question. After detecting all possible classifications, it was necessary to eliminate certain features that, although they had been adopted in these papers, would not be appropriate for our research. The features eliminated at this stage were: features related exclusively to the functioning of a specific social media, features related exclusively to geographical factors in which the research was applied, and features that were not linked to the post. Examples of excluded features are page type (Fan Page / Private Page / Party Page), Twitter post types (Mentions / Direct Messages), photo content features, and features related to the candidate's profile on the SN platform.

Third, to merge duplicated features. Naturally, some features were repeated among the taxonomies, either with the same or similar names, and/or similar meanings. Thus, to avoid redundancy, any duplicated or similar features were merged at this stage.

Fourth, to add new features not contemplated in previous studies. Part of the goal of our research was to analyze features that had not yet been addressed in other investigations or different ways to classify a feature already used in a previous study. Thus, at this stage, we added such features to our taxonomy, and were as follows:

- Aristotelian rhetoric: According to [27], features of texts are related to logos, pathos, or ethos. Logos is an appeal to logic using intellectual reasoning and argument structure such as giving claims, sound reasons for them, and supporting evidence. Pathos is an appeal to the emotions of the audience, often based on claims they hold. By influencing their feelings, the audience may be pushed into taking some action, believing an argument, or responding in a certain way. Ethos is an appeal based on the good character of the author. It involves persuading the audience that the author is credible and well-qualified, or possesses other desirable qualities, which signify that the author's arguments carry weight.
- Rhetorical devices: These are techniques used by an author or speaker to convey meaning to an audience or reader, to persuade them to consider a subject from one perspective, showing a given perspective, or using language designed to encourage or arouse emotion or action. Rhetorical devices evoke emotional responses in the audience using language, but this is not their primary purpose. By doing so, they rather attempt to make a position or argument more persuasive than it would otherwise be.
- Content type: We have defined four types of content types, inspired by [36]. Campaign acts are posts related to publicizing/promoting campaign events, such as rallies, marches, and demonstrations. Personal posts do not contain any political slant, such as posts on sports, or family outings. Political-Ideological posts deal with ideological issues such as minority rights and religion. Political-Purposeful posts deal with propositional issues, such as economy, education, and public health.
- Functional approach: Using the [36] as inspiration, we have defined three functional approaches. Acclamation posts praise another person (political or not) creating a reciprocal relationship between both. In attack posts, the author publicly attacks another person. In defense posts, the author publicly defends another person from an attack.

Fifth: To perform groupings of features that share the same object of analysis. This step was performed with the aim of making our taxonomy more efficient while still ensuring that we did not exclude any relevant features from our analysis. To do this, we made some reductions, where we grouped features that treated the same object of analysis into a single feature with a new label that indicated its new range. In future studies it would be possible not to include these reductions to have an even higher level of detail.

The resulting taxonomy is presented in Table 2.

Table 2: Defined Taxonomy

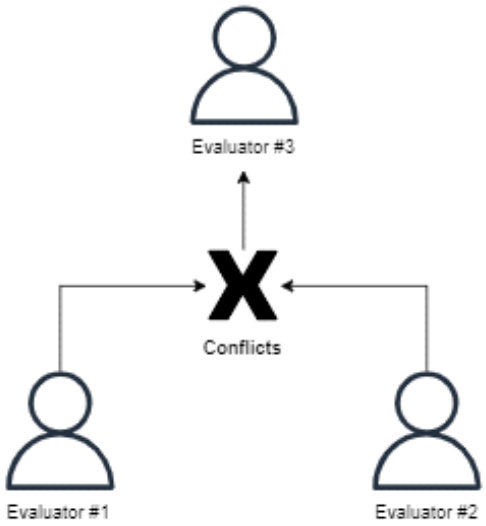
Feature	Description	Possible Classifications
Author	Politician who posted the message	Candidates' names
Time of day	Hour of publication	Post-midnight (12:00AM - 05:59AM); Morning (06:00AM - 11:59AM); Afternoon (12:00PM - 05:59PM); Night (06:PM - 11:59PM)
Day of the week	Day of publishing	Sunday ... Saturday
SM platform	Platform where the post was published	Facebook / Twitter / Instagram
Content structure	Content structure in which the message was written	Photo; Hyperlink; Meme; Text; Video
Aristotelian rhetoric	One of the three classifications of aristotelian rhetoric	Logos; Pathos; Ethos
Rhetorical Device	Rhetorical technique used by the author	Thanks; Collective Appeal; Personal Appeal; Call to Action; Commitment; Political Statement; Endorsement; Fact/statistic; Humor; Opinion; Question; Advertisement; Urgency
Content type	One of the four classifications of content type	Campaign Act; Personal; Political-Ideological; Political-Purposeful
Functional approach	One of the three classifications of functional approach	Acclamation; Attack; Defense

This resulting taxonomy maintains the strengths of the previously proposed taxonomies and improves them with new concepts or even by adding concepts that had not yet been addressed. For example, the taxonomy proposed in contained a feature called "Reasoning" which indicated whether the post used a logical thought to communicate its message. In our proposal, it has been enhanced with more arguments to confirm whether the post "is reasoning or not", such as whether one of the rhetoric used was "Fact/Statistics", whether the type of content of the post is "Political Propositional", or whether the Aristotelian approach adopted was "Logos". Thus, we have introduced more information, and increased our ability to generate new insights.

3.3 Data Classification

The classification of the 1,319 posts was performed manually by 3 different evaluators according to the following process:

Figure 1: Data Classification Process



- *Evaluator #1:* Classify every feature from the taxonomy based on the post content.
- *Evaluator #2:* Classify every feature from the taxonomy based on the post content.
- *Evaluator #3*: Review the classification only for the features that had conflicts between the previous evaluators. In this review, the Evaluator #3 will reclassify the features choosing exclusively between the classifications suggested by the Evaluator #1 and Evaluator #2.

It is worth mentioning that all the evaluators involved in this process classification are also researchers in the context of Politics, and Government Data and therefore have knowledge about the themes and the

context addressed by this research. Furthermore, to ensure better alignment and understanding among the 3 raters and the aspects of this research involved in data classification, training sessions were held with the raters, and an initial set of posts were classified under supervision

3.4 Statistical and descriptive data analysis

Data analysis was performed in two ways: descriptive analysis and statistical analysis.

The descriptive analysis was performed for each feature and its possible classifications, as presented in Table 1. We calculated the total posts, total interactions, minimum and maximum number of interactions, and average and standard deviation. This analysis enables us to highlight a group of features that were reflected in a higher number of interactions per post and, consequently, a higher engagement. It also provided indications of the main differences between the number of interactions received by each type of post.

Although the descriptive analysis allows us to answer the research question, we performed statistical tests on the interactions over features to verify whether eventual differences in engagement found in the descriptive analysis were statistically significant. For this, we used the unpaired t-test (also known as the independent t-test), a statistical procedure that tests two independent or unrelated groups to determine whether there is a significant difference between them [28]. Based on this, we compared posts with different classifications from a specific feature to verify whether there really was a statistically significant difference between them both. According to conventional criteria, a difference is considered statistically significant between the two sets when the *p-value* of the test is lower than *0.05*.

3.5 Data Preparation to Predictive Models

Regarding the prediction experiments, it was necessary to use some data normalization techniques. The first was the One Hot Encoding which is a method of encoding data to prepare it for be used in the algorithm and return the best results. Using the one-hot technique, we set each categorical value into a new categorical column and assign a boolean (1 or 0) to those columns. Each integer value is represented as a binary vector. All the values are zero, and the index is marked with 1. This step was executed inside of the processes to obtain the results and answers regarding the RQ2, described in section 4.2.1, and RQ3, described in section 4.2.2

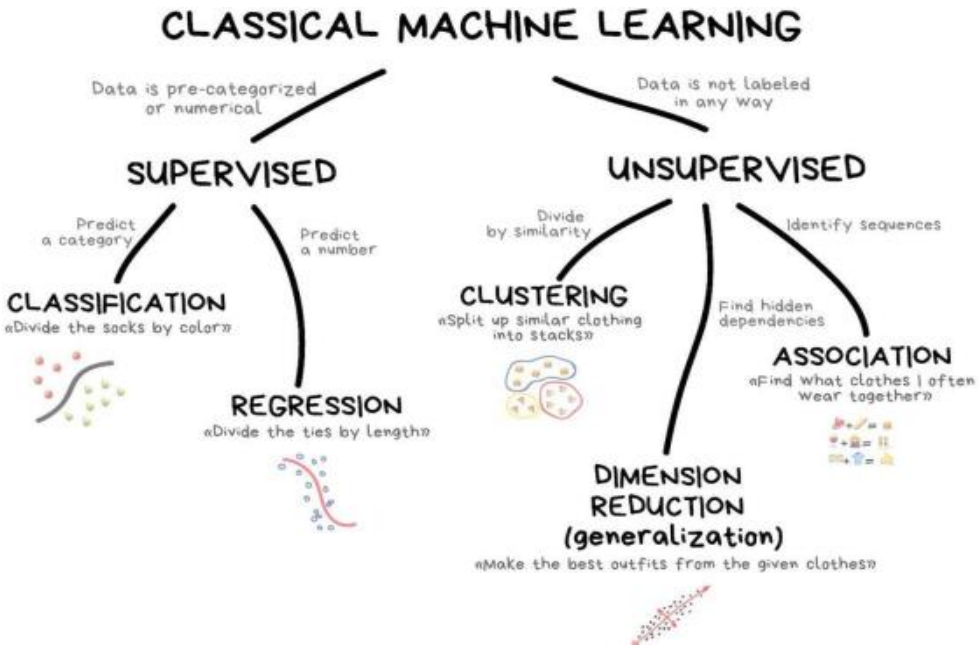
As we propose to build three predictive models, one for general use (with data from all candidates), one specific for the candidate Jair Bolsonaro and another specific for the candidate Fernando Haddad, we needed to add one more step in data preparation for the specific models due to the decrease in the volume of data to be trained in the models, when compared to the generalist model. To deal with that, we use the Principal Component Analysis (PCA). Typically, the central idea involving this technique is to be able to reduce the dimensionality of a dataset involving a large range of interrelated parameters while retaining while much variation of the deviation present in typically the dataset. This is done by simply transforming to a new some of parameters, which will be uncorrelated, and which in turn are ordered and so that the initial few retain almost all the deviation seen in all involving the original changing [39]

3.6 Machine Learning Models to Prediction

Part of the aim of this paper consists of the study of the application of artificial intelligence models to predict the level of engagement of a post, number of likes, given the characteristics present in this post. To this end, we use machine learning models which are mainly characteristic of the fact that given a dataset as input, known as training data, the model can learn from this data, find correlations in it, and predict the behavior of new records that were not initially contained in the input dataset [29]. The literature in general divides the machine learning models in two big areas according to the way that the process of learning will be performed through model. This process can be of two different ways: (i) supervised, when reference of the goal to be achieved is provided, that is, training with the knowledge of the environment. (ii) non-supervision, when no references are used, ie, there is no training with the knowledge of the environment. Our scenario matches with the supervised training model, since we have a training dataset, obtained in the data classification task, and described on section 3.3, which contains the expected results that we aim to achieve

Inside of the supervise training models we also have another division commonly performed in the literature regarding to the type of problem that will be solved, in general with have two types: (i) classification, where it should be used when data submitted to the model wish to predict a label. (ii) regression, which instead of predicting a label, seeks to predict a numerical value [30]. The forecast of the level of engagement of a post on a candidate's social profile is characterized by being a regression problem, as the object of analysis and study is typified as a number [30], the number of likes that the post in question received.

Figure 2: Classical Machine Learning Illustration [31]



That said, we chose to initially use the linear regression model because, it is a descriptive model, that is, the model helps in analyzing the strength of the association between the outcome (dependent variable) and predictor variables [33]; Apart from the linear regression model, we additionally use decision tree -based algorithmic models because they're additionally explainable [30], that is, they allow transparency to the model and help to understand how the predictions were generated [32]. For the decision tree-based algorithmic we choose based DecisionTree, Random Forrest, and Gradient Boosting.

Bellow, you can see more details about each model:

Linear Regression*:* is a statistical procedure for calculating the value of a dependent variable from an independent variable. Linear regression measures the association between two variables. It is a modeling technique where a dependent variable is predicted based on one or more independent variables. Linear regression analysis is the most widely used of all statistical techniques [33]. For this study, we parameterized the Linear Regression model with the following custom settings: fit intercept = true, elastic net regression as regularization based, Regularization strength as Alpha = 0.0001, and Elastic net mixing between L1 and L2 equals to 0.50: 0.5.

DecisionTree: is a commonly used data mining method for establishing classification systems based on multiple covariates or for developing prediction algorithms for a target variable. This method classifies a population into branch-like segments that construct an inverted tree with a root node, internal nodes, and leaf nodes. The algorithm is non-parametric and can efficiently deal with large, complicated datasets without imposing a complicated parametric structure. When the sample size is large enough, study data can be divided into training and validation datasets. Using the training dataset to build a decision tree model and a validation dataset to decide on the appropriate tree size needed to achieve the optimal final model [34] For this study, we parameterized the DecisionTree model with the following custom settings: max_depth = 100, min_samples_leaf = 2, min_samples_split = 5

Random Forrest: are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges a.s. to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them [35]. For this study, we parameterized the DecisionTree model with the following custom settings: number of decision trees in the forest equals to 10 and the number of random features to consider at each split = 5.

Gradient Boosting: is based in a learning procedure consecutively fits new versions to supply a more accurate estimate of the response variable. The principle thought behind this algorithm is to build the newly created base-learners to be maximally correlated with the negative gradient of the loss function, associated with the complete ensemble [38]. For this study, we parameterized the Gradient Boosting model with the

following custom settings: learning_rate = 0,100, n_estimators =100, for individuals tree the max_depth = 3, min_samples_split = 2, and fraction of training instance equals to 1,00.

Figure 3: Linear Regression Algorithm

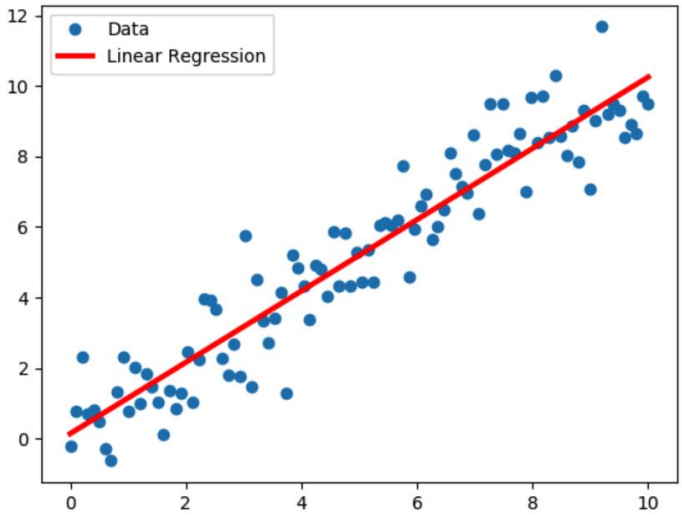


Figure 4: DecisionTree Algorithm

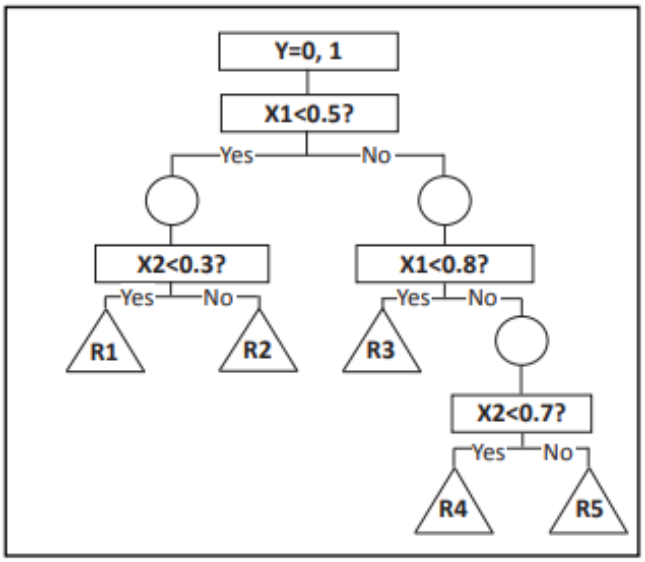


Figure 5: Random Forest Algorithm

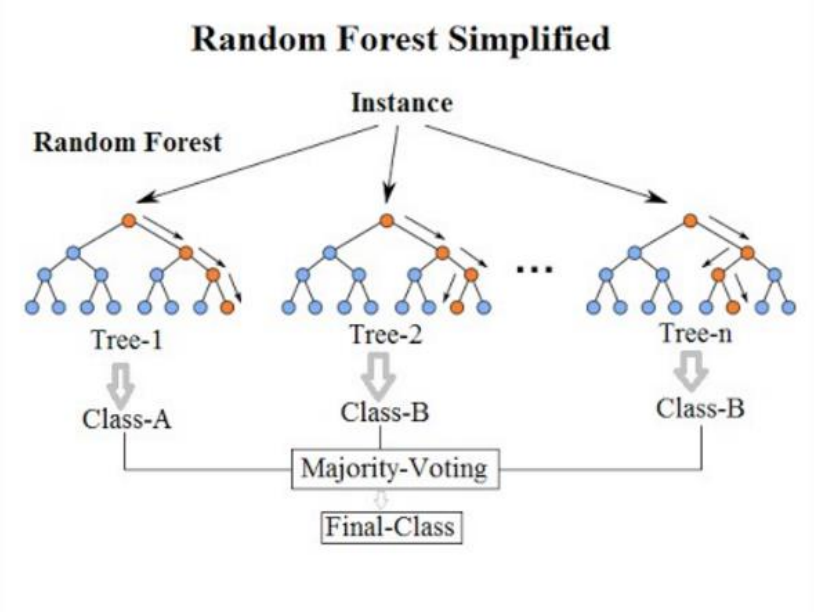
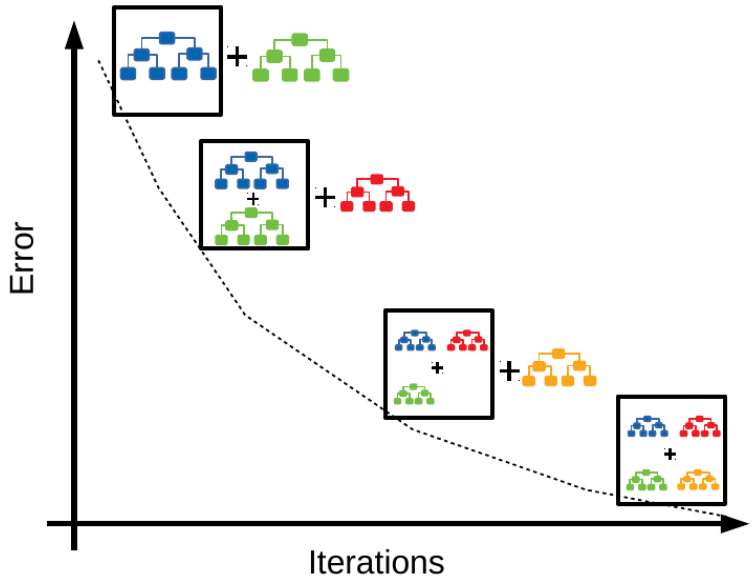


Figure 6: Gradient Boosting Algorithm



When carrying out the experiments with the aforementioned algorithms, no customizations were made in the specific parameterization of each algorithm. Because, since algorithms based on decision trees were used, optimal hyperparameters are recommended for the models and not to act in a way [37]. Regarding the division of data between training and testing, we used the ratio of 60/40 randomly.

3.7.1 Evaluation of Machine Learning Prediction Models

To analyze the level of accuracy of the models regarding the prediction of engagement of a post, we use the following metrics as study parameter

- *R-squared:* represents the proportion of variance in the dependent variable that is explained by the linear regression model. It is an unscaled score, that is, regardless of whether the values are small or large, the value of R squared will be less than one.

- *Root Mean Squared Error (RSME):* is the square root of the Mean Squared error. It measures the standard deviation of residuals.

- *Mean absolute error (MAE):* represents the mean of the absolute difference between the actual and predicted values in the data set. It measures the average of the residuals in the data set.

4. RESULTS

According to the methodology, we analyzed 1,319 posts made by the 13 candidates on Facebook, Twitter, and Instagram for the 2018 Brazilian presidential elections over a two-week period prior to election day. As also detailed, we carried out experiments using statistical tools, for descriptive and statistical analysis, and machine learning algorithms to answer the research questions earlier described in section 3.0.

To support the answer to *RQ1* we will use the data obtained through the results of descriptive and statistical analyses, you can see those results at section 4.1.

The support from the answers to the *RQ2* and *RQ3* are found in section 4.2 once, those answers were obtained with the prediction models.

4.1 Data Analysis: Descriptive and Statistical

Tables 3 to 11 present the descriptive data. For ease of viewing, the tables have been categorized into descending order by the average number of interactions.

Considering all candidates and all SM together, Table 2 demonstrates that Jair Bolsonaro was by far, the candidate with the highest absolute and average number of interactions on his posts, 14.2 million and 167,386 interactions respectively. In second place was Ciro Gomes, who obtained just 3.5 million total interactions and an average of 20,968 interactions per post during the same period. It is important to highlight that from the beginning of his campaign, Jair Bolsonaro focused his political marketing strategy on digital media, either by strategic factors or limitations imposed on him, as for example, the extremely limited TV time compared to the other candidates. This fact illustrates the power of social media in the political environment as a political marketing tool to convert digital engagement and interaction into votes in the election.

It is also important to pay attention to the moment when the post is published, to ensure the targeting of the message at the right moment to obtain more visualization and engagement. As presented in Table 3, users reacted more on posts made in the afternoon (Average = 30,665) and less post-midnight (Average = 10,937) and this difference was proven to be statistically significant ($p < 0.05$).

Regarding the day of week, posts made on Friday (Average = 29,342), Thursday (Average = 26,501) and Saturday (Average = 25,888) were on average more likely to engage, with no significant statistical difference between them. On the other hand, posts made on Wednesday (Average = 11,294) were less likely to engage. The difference between the engagement of posts made on the day's most likely to engage and posts made on Wednesday are statistically significant. This conclusion is interesting since only 7% of posts were made between Thursday and Saturday, but it is the interval with most engagement.

Table 3: Engagement by Author

Author	Posts	Interactions				
		Minimum	Maximum	Std Dev	Sum	Average
Jair Bolsonaro	85	2.645	1.089.320	233.470	14.227.821	167.386
Ciro Gomes	166	-	145.985	27.151	3.480.623	20.968
Cabo Daciolo	42	-	99.743	25.124	578.094	13.764
João Amoêdo	210	-	150.090	22.721	2.482.927	11.823
Fernando Haddad	194	-	129.063	18.419	2.123.319	10.945
Guilherme Boulos	152	-	125.978	16.421	1.160.524	7.635
Marina Silva	77	-	130.191	16.065	549.422	7.135
Henrique Meirelles	121	-	41.809	5.917	202.777	1.676
Geraldo Alckmin	63	-	6.732	1.574	65.043	1.032
Alvaro Dias	160	-	21.110	2.680	160.796	1.005
Vera Lúcia	6	28	1.905	806	4.283	714
Eymael	16	30	2.579	815	8.599	537
João Goulart Filho	27	33	3.311	623	6.682	247

Table 4: Engagement by Time of Day

Time of day	Posts	Interactions				
		Minimum	Maximum	Std Dev	Sum	Average
Afternoon	366	-	1,089,320	104,951	11,223,508	30,665
Morning	396	-	949,411	64,452	6,300,480	15,910
Night	446	-	778,794	49,318	6,312,914	14,155
Post-midnight	111	-	392,515	42,984	1,214,008	10,937

Table 5: Engagement by Day of the week

Day of the week	Posts	Interactions				
		Minimum	Maximum	Std Dev	Sum	Average
Friday	18	620	121.728	31.278	528.150	29.342
Thursday	8	4.948	63.338	23.263	212.007	26.501
Saturday	62	13	149.935	37.419	1.605.028	25.888
Monday	841	-	949.411	80.458	17.450.678	20.750
Tuesday	281	-	1.089.320	69.880	3.929.729	13.985
Sunday	37	-	125.978	28.665	512.173	13.843
Wednesday	72	1	155.323	28.455	813.145	11.294

One of the differentials of this research is the fact that we obtained and analyzed data from the three main social medias in use in Brazil: Facebook, Instagram, and Twitter. This is important because candidates very often post the same content on the three medias simultaneously. Results are presented in Table 6 and are in line with the results of [5], showing that on overage Instagram was the media that most engaged (Average = 42,913). The difference in engagement between Instagram and the other platforms was shown as statically significant.

Table 6: Engagement by Social Media platform

Social Network Platform	Posts	Interactions				
		Minimum	Maximum	Std Dev	Sum	Average
Instagram	255	166	949.411	113.549	10.942.798	42.913
Facebook	425	5	1.089.320	88.755	11.492.800	27.042
Twitter	639	-	99.743	11.734	2.615.312	4.093

Regarding content structure, *memes* obtained the highest average (42,670) of interactions, but there were very few posts, insufficient to be statistically significant. The same occurred with hyperlinks. Posts purely containing text obtained a lower average of interactions (11,973). This result is expected because photos and videos are generally considered to be more appealing to users.

Table 7: Engagement by Content Structure

Content Structure	Posts	Interactions				
		Minimum	Maximum	Std Dev	Sum	Average
Meme	18	-	150.485	53.318	682.713	42.670
Hyperlink	92	-	1.089.320	160.946	3.149.775	36.625
Photo	377	-	949.411	89.130	10.048.251	28.627
Video	425	-	842.939	53.946	5.270.728	13.310
Text	407	-	538.829	41.445	4.549.582	11.973

In relation to Aristotelian rhetoric, the results presented in Table 8 demonstrate that rhetoric based on pathos, i.e., appealing to the audience's emotions, usually based on claims they already hold, was the one that generated the most interactions from users. On the other hand, posts based on logos, i.e., based on logic using intellectual reasoning, received the lower average of interactions. The superiority of engagement on pathos posts presented a statistically significant difference when comparing both with ethos, ($p = 0.02561$), the second rhetoric that engages more, and logos ($p < 0.005$), the rhetoric with the lowest engagement.

Table 8: Engagement by Aristotelian rhetoric

Aristotle's rhetoric	Posts	Interactions				
		Minimum	Maximum	Std Dev	Sum	Average
Pathos	443	-	949.411	118.694	61.498	61.498
Ethos	749	-	1.089.320	58.237	83.288	21.972
Logos	127	-	463.630	76.176	50.340	13.218

Table 9 presents the results of the analysis of the rhetorical device adopted. Humorous posts (Average = 151,344) tended to engage much more than other rhetoric's, the same as indicated in [3], followed by political statements. On the other hand, the rhetoric's least likely to engage users are questions (Average = 1,451), personal appeal (Average = 1,636), and sense of urgency (Average = 2,879). When applying the statistical tests comparing the two most engaging rhetoric's and the three least engaging, the difference between the two sets is statistically significant. However, it is important to highlight the low number of samples regarding these devices, indicating that these results may be biased.

Table 9: Engagement by Rhetorical Device

Rethorical Device	Posts	Interactions				
		Minimum	Maximum	Std Dev	Sum	Average
Humor	9	-	949.411	302.426	1.362.100	151.344
Political Statement	102	-	842.939	96.635	3.194.415	31.318
Fact/statistic	41	-	463.630	76.121	1.029.010	25.098
Endorsement	56	-	778.794	105.933	1.311.822	23.425
Advertisement	107	-	828.522	98.736	2.253.895	21.064
Collective Appeal	216	-	538.829	52.007	4.062.189	18.806
Call to Action	317	-	1.089.320	63.770	5.379.014	16.794
Opinion	109	-	243.389	38.594	1.771.519	16.252
Commitment	120	1	664.945	63.238	1.860.666	15.506
Other	89	-	157.956	31.605	1.291.920	14.516
Thanks	121	-	493.784	53.105	1.459.306	12.060
Urgency	19	-	16.014	4.549	54.708	2.879
Personal Appeal	8	2	7.883	2.652	13.090	1.636
Question	5	2	6.836	3.014	7.256	1.451

Results regarding the functional approach are presented in Table 10. On this aspect, the posts with the defense approach were, on average, those with more engagement (Average = 51,236), followed by the posts with the attack approach (Average = 22,689). Thus, on average, the posts with the least engagement were the acclamation posts. Based on the statistical tests, it is possible to state that the defense posts had a statistically higher engagement ($P < 0.0001$) than the acclaim posts, despite the low number of posts.

Table 10: Engagement by Functional Approach

Functional Approach	Posts	Interactions				
		Minimum	Maximum	Std Dev	Sum	Average
Defense	18	1	392.515	118.694	922.254	51.236
Attack	195	-	538.829	58.373	4.424.285	22.689
Acclamation	995	-	1.089.320	77.171	17.892.133	17.982
Other	111	-	290.617	39.858	1.812.238	16.326

Lastly, Table 11 presents the results regarding the type of content. The purposeful political posts were the ones that, on average, promoted the least reaction from users (Average = 8,035), however, the personal posts were the ones that generated the most engagement (Average = 27,939), followed by political ideology (Average =20,445). When comparing ideologically driven and purposeful driven posts, it is possible to observe a statistically significant difference in the level of engagement (p = 0.0016), as well as when comparing personally driven and purposeful driven posts (p = 0.0319).

Table 11: Engagement by Content Type

Content Type	Posts	Interactions				
		Minimum	Maximum	Std Dev	Sum	Average
Personal	235	-	949.411	108.220	6.565.598	27.939
Political-Ideological	243	-	349.229	43.763	4.968.142	20.445
Campaign Act	614	-	1.089.320	77.143	11.163.377	18.181
Other	90	-	157.956	31.411	1.252.952	13.922
Political-Purposeful	137	-	150.485	17.116	1.100.841	8.035

4.2 Engagement Prediction Results

Regarding the prediction step, we can separate into two different groups, according to our research questions described in section 3.0.

4.2.1 General Prediction Results

The first group of results corresponds to RQ1, which refers to the proposal to use machine learning models to predict the level of engagement of posts made by politicians in their social networks. For that we obtain the following results:

Table11: Results of Posts Predictions

Score Results			
Model	**RMSE**	**MAE**	**R2**
Gradient Boosting	25660	9259	0.77
Random Forest	26283	8264	0.76
DecisionTree	27027	8842	0.75
Linear Regression	44952	20203	0.30

As we can see the model with the best performance was the Gradient Boosting (R2 = 0.77), followed by Random Forest (R2 = 0.76), DecisionTree (r2= 0.75), and Linear Regression (r2 = 0.30) which had the worst performance among the selected models.

It was also possible to identify the most relevant features, considering the R2 metric, obtained after training the prediction for each of the selected models. You can see the top 10 features, in importance considering the r2 metric, from each model below:

Table 12: Results of General Feature Importance

Feature Importance

Linear Regression			DecisionTree		
Order	Feature	Score	Order	Feature	Score
1	SM=instagram	7229.99	1	Author=Jair Bolsonaro	7141.07
2	Author=Jair Bolsonaro	7037.95	2	SM=instagram	5637.98
3	SM=twitter	6043.34	3	SM=facebook	4925.08
4	Content Structure=photo	3445.88	4	Rhetorical Device=humor	2067.65
5	Content Type=personal	3270.16	5	Functional Approach=defense	1928.56
6	Rhetorical Device=political statement	2489.41	6	Content Type=personal	1698.4
7	Author=Alvaro Dias	1891.09	7	Content Structure=photo	1621.91
8	Author=Guilherme Boulos	1847.7	8	Rhetorical Device=opinion	1508.31
9	Rhetorical Device=commitment	1763.22	9	Rhetorical Device=political statement	1476.47
10	Content Type=political purposeful	1568.72	10	Time of Day=night	1.461

Random Forrest			Gradient Boosting		
Order	Random Forrest	Score	Order	Gradient Boosting	Score
1	Author=Jair Bolsonaro	15486.3	1	Author=Jair Bolsonaro	13913.6
2	SM=instagram	9790.55	2	SM=instagram	9857.63
3	SM=twitter	4149.67	3	SM=twitter	4134.4
4	Author=João Amoêdo	2404.26	4	Rhetorical Device=political statement	1733.64
5	Author=Ciro Gomes	2114.5	5	Author=Ciro Gomes	1654.01
6	Rhetorical Device=political statement	1824.49	6	Author=João Amoêdo	1628.93
7	Author=Fernando Haddad	1676.09	7	SM=facebook	1100.38
8	SM=facebook	1178.31	8	Day of the Week=saturday	1093.76
9	Day of the Week=saturday	1110.56	9	Content Type=personal	927.758
10	Content Type=personal	1026.51	10	Time of Day =afternoon	853.396

4.2.1 Candidate Prediction Results

The second group of results concerns about RQ3, which aims to id/entity differences in the way the first 2 candidates, ranked by electoral results, and from opposing political sides, direct their communications to citizens on their social media. When executing the machine learning models just with the data regarding to Jair Bolsonaro, we have the following results:

Table 13: Results of Jair Bolsonaro Posts Predictions

Score Results			
Model	RMSE	MAE	R2
Decision Tree	81439	45254	0.77
Gradient Boosting	90044	51663	0.72
Random Forest	94551	50861	0.69
Linear Regression	132816	94426	0.39

The Decision Tree was the model with better results (R2= 0.77), followed by Gradient Boosting (R2 =0.72), Random Forest (R2=0.69), and Linear Regression (R2 = 0.39). It is worth noting that the DecisionTree model performed better not only considering R2 but also the other metrics (MAE = 45254 and RMSE = 81439). It is also worth making some considerations about these results. For example, considering that the candidate in question, Jair Bolsonaro, has a mean number of interactions equal to 167.386 (Table 3) and a standard deviation equal to 233.470 (Table 3), quite high, we can say that the MAE (Table 13) obtained is the DecisionModel is a good result. Taking all this into account we can say that the result obtained was satisfactory.

In this experiment was also possible to identify the most relevant features, considering the R2 metric, obtained after training the prediction using just data from Jair Bolsonaro for each of the selected models. You can see the top 10 features, in importance considering the r2 metric, from each model below:

Table 14: Jair Bolsonaro Feature Importance

Feature Importance									
	Linear Regression				DecisionTree				
Order	Feature	Score			Order	Feature			Score
1	SM=twitter	52589.9			1	SM=instagram			71011.3
2	SM=instagram	45.828			2	SM=facebook			39.794
3	Rhetorical Device=political statement	27760.7			3	Rhetorical Device=political statement			12082.7
4	Content Type=personal	22354.1			4	Content Type=personal			11040.8
5	Time of Day=afternoon	19.681			5	Functional Approach=acclamation			10645.4
6	Time of Day=morning	19465.9			6	Content Structure=text			10063.5
7	Content Structure=text	19149.6			7	Content Structure=video			4806.71
8	Content Structure=video	16624.7			8	Rhetorical Device=opinion			3891.24
9	Rhetorical Device=fact/statistic	13053.7			9	Time of Day=Morning			3408.83
10	Time of Day=night	11608.6			10	Content Structure=photo			3123.98
	Random Forrest					Gradient Boosting			
Order	Feature	Score			Order	Feature			Score
1	SM=instagram	101.198			1	SM=instagram			100966
2	SM=twitter	31672.8			2	Rhetorical Device=political statement			25385.5
3	Rhetorical Device= political statement	26248.3			3	Content Type=personal			20639.2
4	SM=facebook	12972.1			4	SM=facebook			20417.1
5	Content Type=personal	8237.25			5	SM=twitter			18534.6
6	Content Structure=text	7658.3			6	Content Structure=video			8362.24
7	Functional Approach=attack	7244.08			7	Time of Day=afternoon			8311
8	Time of Day=morning	4792.82			8	Time of Day=morning			6566.23
9	Rhetorical Device=opinion	4268.74			9	Functional Approach=attack			6560.47
10	Functional Approach=acclamation	2782.92			10	Functional Approach=acclamation			1592580

When executing the machine learning models just with the data regarding to Fernando Haddad, we have the following results:

Table 15: Results of Fernando Haddad Posts Predictions

Score Results			
Model	RMSE	MAE	R2
Decision Tree	8997	4562	0.58
Random Forest	9840	4919	0.50
Gradient Boosting	10363	5405	0.45
Linear Regression	11713	6923	0.29

The Decision Tree was the model with the better result (R2= 0.58), followed by Random Forest (R2 =0.50), Gradient Boosting (R2=0.45), and Linear Regression (R2 = 0.29). However, there was a drop in the general performance of the models, considering the results obtained in the previous experiments (Table 11 and Table 13). However, considering that the candidate in question, Fernando Haddad, obtained an average of interactions in his posts equal to 10.945 and a standard deviation equal to 18.419 (Table 3), the results obtained by DecisionTree with MAE = 4.562 e RMS = is still within the acceptable range.

The top 10 relevant features, considering the R2 metric, obtained after training the prediction using just data from Fernando Haddad for each of the selected models is below:

Feature Importance						
	Linear Regression				DecisionTree	
Order	Feature	Score		Order	Feature	Score
1	SM=instagram	4353.35		1	SM=instagram	5422.18
2	SM=twitter	3.349		2	SM=facebook	3279.82
3	Rhetorical Device=commitment	1986.45		3	Content Structure=video	2097.68
4	Content Structure=video	1411.73		4	Time of Day=night	1546.92
5	Time of Day=night	1119.25		5	Day of the Week=saturday	1161.05
6	Day of the Week=saturday	1064.1		6	Time of Day=morning	761.925
7	Content Structure=text	916.856		7	Day of the Week=monday	704.524
8	Rhetorical Device=collective appeal	881.612		8	Rhetorical Device=collective appeal	690.765
9	Content Structure=photo	875.591		9	Content Type=campaign act	564.782
10	SM=facebook	846.369		10	Rhetorical Device=commitment	388.464
	Random Forrest				Gradient Boosting	
Order	Feature	Score		Order	Feature	Score
1	SM=twitter	5415.34		1	SM=instagram	6016.53
2	SM=instagram	3937.46		2	SM=twitter	2196.66
3	Day of the Week=saturday	1724.12		3	Rhetorical Device=collective appeal	2069.35
4	Time of Day=night	1486.24		4	Time of Day=night	1850.77
5	Rhetorical Device=collective appeal	1450.56		5	Content Structure=video	1848.09
6	Content Structure=video	1135.16		6	Day of the Week=saturday	1462.82
7	SM=facebook	538.22		7	SM=facebook	1073.56
8	Rhetorical Device=fact/statistic	243.96		8	Day of the Week=monday	679.502
9	Content Type=campaign act	233.081		9	Content Type=political purposeful	387.84
10	Content Structure=text	230.251		10	Rhetorical Device=urgency	375.609

Table 16: Fernando Haddad Featured Importance

As the DecisionTree model got the best prediction results for both candidates, we will use feature importance of it to do a more detailed analysis to find similar or differences between the digital strategies from the two candidates. You can find the results of this analysis for each feature bellow:

Regarding the time of day when the post was published, we did not have many differences between the candidates, except for the fact that for the left candidate publishing the post in the evening impacts considerably the level of his engagement when compared to the other times, and for the right candidate the impact is greater when the posts are published during the morning or afternoon, in that order. A similarity is that for the time of the day that has the least impact is post-midnight.

On the day of the week, Monday and Saturday lead the rankings for impact on engagement compared to the other days of the week for both candidates. Furthermore, we find no substantial differences or similarities

Regarding social media, the order of impact on engagement was the same for both candidates, being Instagram, Facebook, and Twitter.

In relation to the content structure, there were also not many differences, the only difference consists of the swapping of positions between 1st and 2nd in the content structure which were inverted between the left candidate, in which text type leads followed by video, and the right candidate, in which video leads followed by text. In both candidates, the list follows with photo and hyperlink in last place

About Aristotle's Rhetoric, on the right candidate, the Pathos rhetoric had the most impact while the Ethos rhetoric had the most impact on the left candidate. Another important observation is that no posts classified as Logos were found on the right candidate while on the left candidate this rhetoric is the second most impacting on engagement.

Regarding the rhetorical device, we had some differences between the two candidates, for example: for the right-wing candidate, the devices that most impacted were political statement and personal appeal. When compared to the left candidate, the personal appeal was one of the devices that least impacted the engagement, second only to urgency, on the other hand, collective appeal led this rank to the left-wing candidate followed by commitment and call to action, this one being one of the least impacted on the engagement of the right-wing candidate's posts. However, there is a similarity in both, and it consists of the little impact of the presence of the fact/statistic device on the engagement of the posts of both.

About the content type, for the right candidate we had the following rank, from the highest impact to the lowest: personal, campaign action, and political-ideological, it is worth noting the absence of the political purposeful type. For the left candidate, the rank follows with campaign act, political purposeful, personal, and political ideological. We can see that the campaign act stays between the top of the rank for both candidates

Finally, regarding the Functional Approach we had the same order for both candidates, that was: acclamation, attack, and defense.

4.3 Summary and Discussion of results

The findings observed in the analysis together with their impacts may be summarized as:

RQ1. Which features of social media posts have the greatest influence over user engagement on the social media profiles of candidates in the run-up to elections?

- Instagram was the social media with the most interactions among the posts by users and politicians. This information is in accordance with the results found in the paper [12] and is significant, since, with this knowledge, it is possible to take advantage of it by using the Instagram channel, even more, to generate engagement and strengthen the relationship between users and politicians or also create strategies to raise the level of interaction on other social media.
- It was also possible to evidence that posts made in the afternoon usually received more engagement. Posts made from Thursday to Saturday were more likely to engage, even though campaigns focused on posts on Mondays and Tuesdays. Thus, results have suggested that campaigns should focus on days with the most engagement or change the strategy to other days.
- It was confirmed that posts with memes, photos, and videos were more appealing than just textual content.
- In accordance with [3], humorous posts, although there were only a few samples, received more engagement than other kinds of posts.

- It was discovered that Brazilian voters represented in social media tend to engage more when exposed to emotional discourse (pathos), than to more rational discourses (logos).
- The analysis confirmed the polarized scenario, whereby (mostly) defensive and offensive posts received more interactions than others.
- Lastly, it was found that personal posts, unrelated to the campaign, demonstrated the highest level of engagement. On the other hand, political-purposeful posts received a low number of interactions. The results suggest that voters on SM may be more interested in content that represents candidates as "normal people" than in content providing political proposals.

RQ2. Is it possible to use Machine Learning models to predict the engagement of politicians' social media posts in the run-up to elections?

- Yes, it is possible to predict the engagement of politicians' social media using the Gradient Boosting machine learning model, with R2= 0.77 (Table 11). In addition to Gradient Boosting Forest, other models were tested, namely: Linear Regression, DecisionTree, and Random Forest, but they obtained lower performance.
- When the aim is to predict the engagement of one specific politician, the Decision Tree is the best option, because in both experiments done, (Table 13 and Table 15), the model obtained better results compared with the other models.

RQ3. Considering the Brazilian political scenario in the 2018 elections, it is possible to identify differences in digital strategies, between the two candidates with the best electoral results and from opposite ideological political spectrums?

- About the similarities between both candidates, we found the following points. Regarding the day of the week, the days that most impact the engagement of the posts for both are Monday and Saturday. About social media, we had the same order of impact for both candidates, being it from the SM that most impacts to the one that least impacts engagement: Instagram, Facebook, and Twitter. Regarding the content structure, we also had a lot of similar points between both candidates, the only difference consists of the swapping of positions between 1st and 2nd place in which for the left candidate text type leads followed by video, and for the right candidate, the video leads the rank followed by text. In relation to the functional approach, we had no difference in the order of functional approach types according to the level of impact for both candidates, the order was: acclamation, attack, and defense. For the time of the day that the post was published, we found that for both candidates the time of the day that has the least impact is post-midnight. We also found that posts classified as personal and campaign act, in the content type, have more impact on both candidates when compared to the other content types. And finally, another similarity consists of the little impact of the presence of the fact/statistic device on the engagement of the posts for both candidates.
- About the main differences between both candidates, we found the points below. The first big difference is regarding Aristotle's rhetoric which consists in the fact that the Pathos rhetoric had the most impact on the right candidate while the Ethos rhetoric had the most impact for the left candidate. Still, about Aristotle's rhetoric, another big change is that no posts classified as Logos were found on the right candidate while on the left candidate this rhetoric is the second most impacting on engagement. Regarding the rhetorical device, we had some differences between the two candidates, for example, the devices that most impacted were political statement and personal appeal, for the right-wing candidate, and for the left-wing, collective appeal led this rank to the left-wing candidate followed by commitment and call to action. Still about the rhetorical devices another comparison that can be done is for the right-wing, the collect appeal is one that impacts more for the right-wing candidate and that has less impact for the left-wing candidate. The same happened with the call to action which for the left-candidate it is one that has more impact on the level of engagement and for the right candidate is one of less impact. And finally, the last difference is regarding the content type, for the right-wing candidate the presence or absence of the content type politic-ideologic has a huge impact on the level of engagement, which contrasts with the left-wing candidate that has the politic-purposeful with one of the rhetorical devices that most impact his posts engagements

5. THREATS TO VALIDITY

In this section, we will point out the possible threats to validity in this study, despite the care and precautions taken during all stages of the study.

5.1 Generalization

One threat to the validity of this study is the fact that it was based on a single context, the 2018 Brazilian presidential elections. Thus, the results obtained in this research cannot be generalized to other contexts.

5.2 Data Classification

Despite the rigid methodology applied in the data classification process, which relied on three evaluators, all with knowledge and background on the subject and research, submitted to training and supervision sessions, even so, this process may have contained misinterpretations of the posts and consequently errors in the classification of the features.

5.3 Machine Learning Parameters Selection

It was not the purpose of this study to extend the study regarding the selection and customization of the parameters of the machine learning models applied in the prediction process, so we chose to use the default parameter settings of each model. This choice may reflect in a loss of performance in the models and therefore is characterized as a possible threat to validity.

6. CONCLUSIONS

This paper has presented a study on the relationship between posts published on politicians' profiles and their level of engagement during the campaign for the Brazilian presidential election in 2018. To discover the correlation between the features of the posts and the level of engagement we used metrics from the three major social media platforms used during the campaign: Facebook, Twitter, and Instagram. For this, we collected data (N = 1,319) on posts from candidates within these media during the final two weeks before the first round of elections. We also proposed a new taxonomy for the classification of posts, with the aim of taking advantage of the strong points of the taxonomies published in previous studies, but also to innovate by including new features to be investigated and indicated as possible factors that lead to greater interaction of the posts by the users. Hence, posts were classified according to the defined taxonomy.

With the previously classified Facebook, Instagram, and Twitter posts, we performed a descriptive analysis in order to deepen our understanding of the data obtained, and then performed a statistical analysis using the unpaired t-test to assert whether the presence or absence of a certain characteristic in the post was influential to the level of negation and interaction of users with the post in question. After these analyses it was possible to conclude that: Instagram is the social media that engages the most. Posts made in the afternoon and from Thursday to Saturday were more likely to engage, in opposition to posts on Wednesdays. Posts with memes, photos, and videos, rather than just textual content, were more likely to engage, along with humorous posts. Polarized discourses, presenting defense and attack speeches, were more likely to engage. And finally, emotionally charged topics tended to engage more than posts based on rational discourses. Moreover, personal posts, unrelated in any way to politics, engaged more than posts related to political proposals.

We also tested different machine learning models to predict the level of engagement and it was possible to conclude that Gradient Boosting is the best model, among those selected in this paper, to predict the level of engagement among the overall political candidate's social media. However, when the goal is to analyze and predict the level of engagement of a specific political candidate, the DecisionTree model is the best option, when compared to the other models analyzed in this study.

Finally, the third objective of this work was to analyze similarities and differences between the digital strategies and their impacts on the two candidates with better electoral results and from opposite political sides, regarding that we found that although the two candidates are from opposite political facets, both digital strategies had a lot of similarities when it comes to the factors that influence the engagement or not of their posts. For example, the day of the week that the post was published with more impact, Monday, and Saturday, as well as the order of social media, Instagram, Facebook, and Twitter that most impacted the level of engagement were the same. The same happens with the content structure that contains text type and video

type in the top two rankings of the structure of the content that most impacts the engagement of the two candidates. And, again we can find the same behavior with the content type feature because posts classified as personal, and campaign are in the top two rankings of the types of content that most impact the engagement of the two candidates. Despite so many similarities, we also find some differences, which will be reported below. For example, we have a difference in Aristotle's rhetoric that most impacts the engagement of the candidates, Ethos for the left-wing candidate and Pathos for the right candidate, and the third Aristotle's rhetorical that most impacts the left-wing candidate, Logos, it is not even presented in the posts analyzed from the right candidate. About the rhetorical device, has more impact on the left-wing candidate's engagement to use the collective appeal and for the right-wing candidate, the personal appeal has more impact. The same behavior happens with the content type, the politic-ideologic type has a huge impact on the level of engagement for the right-candidate, and for the left-wing candidate that the politic-purposeful is the one that most impact his posts engagement.

In terms of future studies, there are some points that should be further explored. First, the analysis was focused on the two weeks just before the elections. Thus, longer time intervals should be analyzed, to find changes in user behavior during the campaign. Periods outside the electoral campaign could also be analyzed, to search for similarities and differences regarding non-electoral periods. Regarding the prediction, it is possible to explore other models or even customization the previous models selected. Moreover, applying the same taxonomy and methodology of this study in other countries and other elections may help to understand similarities and differences that may occur due to geo-political factors

REFERENCES

[1] A. Jungherr, "Twitter use in election campaigns: A systematic literature review," *Journal of Information Technology and Politics*. 2016, doi: 10.1080/19331681.2015.1132401.

[2] R. Heiss, J. Knoll, and J. Matthes, "Pathways to political (dis-)engagement: motivations behind social media use and the role of incidental and intentional exposure modes in adolescents' political engagement," *Communications*, vol. 45, no. s1, pp. 671–693, Nov. 2020, doi: 10.1515/commun-2019-2054.

[3] R. Heiss, D. Schmuck, and J. Matthes, "What drives interaction in political actors' Facebook posts? Profile and content predictors of user engagement and political actors' reactions," *Information, Commun. Soc.*, vol. 22, no. 10, pp. 1497–1513, Aug. 2019, doi: 10.1080/1369118X.2018.1445273.

[4] Pew Research Center, "2016 Campaign: Strong Interest, Widespread Dissatisfaction," 2016. [Online]. Available: https:/www.pewresearch.org/politics/2016/07/07/2016-campaign-strong-interest-widespread-dissatisfaction/.

[5] K. dos S. Brito, N. Paula, M. Fernandes, and S. Meira, "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*, Jun. 2019, pp. 332–341, doi: 10.1145/3325112.3325252.

[6] O. Falck, R. Gold, and S. Heblich, "E-lections: Voting Behavior and the Internet," *Am. Econ. Rev.*, vol. 104, no. 7, pp. 2238–2265, Jul. 2014, doi: 10.1257/aer.104.7.2238.

[7] F. Campante, R. Durante, and F. Sobbrio, "Politics 2.0: The Multifaceted Effect of Broadband Internet on Political Participation," *J. Eur. Econ. Assoc.*, vol. 16, no. 4, pp. 1094–1136, Aug. 2018, doi: 10.1093/jeea/jvx044.

[8] R. Enikolopov, A. Makarin, and M. Petrova, "Social Media and Protest Participation: Evidence From Russia," *Econometrica*, vol. 88, no. 4, pp. 1479–1514, 2020, doi: 10.3982/ECTA14281.

[9] M. Manacorda and A. Tesei, "Liberation Technology: Mobile Phones and Political Mobilization in Africa," *Econometrica*, vol. 88, no. 2, pp. 533–567, 2020, doi: 10.3982/ECTA14392.

[10] K. dos S. Brito, R. L. C. Silva Filho, and P. J. L. Adeodato, "A Systematic Review of Predicting Elections Based on Social Media Data: Research Challenges and Future Directions," *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 4, pp. 819–843, Aug. 2021, doi: 10.1109/TCSS.2021.3063660.

[11] J. S. Santos, F. Bernardini, and A. Paes, "A survey on the use of data and opinion mining in social media to political electoral outcomes prediction," *Soc. Netw. Anal. Min.*, vol. 11, no. 1, p. 103, Dec. 2021, doi: 10.1007/s13278-021-00813-4.

[12] K. dos S. Brito, S. R. de L. Meira, and P. J. L. Adeodato, "Correlations of social media performance and electoral results in Brazilian presidential elections," *Inf. Polity*, pp. 1–23, Jun. 2021, doi: 10.3233/IP-210315.

[13] B. Bimber, "Digital Media in the Obama Campaigns of 2008 and 2012: Adaptation to the Personalized Political Communication Environment," *J. Inf. Technol. Polit.*, vol. 11, no. 2, pp. 130–150, Apr. 2014, doi: 10.1080/19331681.2014.895691.

[14] D. L. Cogburn and F. K. Espinoza-Vasquez, "From Mediaed Nominee to Mediaed Nation: Examining the Impact of Web 2.0 and Social Media on Political Participation and Civic Engagement in the 2008 Obama Campaign," *J. Polit. Mark.*, vol. 10, no. 1–2, pp. 189–213, Feb. 2011, doi: 10.1080/15377857.2011.540224.

[15] N.-A. Hall, “Understanding Brexit on Facebook: Developing Close-up, Qualitative Methodologies for Social Media Research,” *Sociol. Res. Online*, p. 17, Oct. 2021, doi: 10.1177/13607804211037356.
[16] P. L. Francia, “Free Media and Twitter in the 2016 Presidential Election: The Unconventional Campaign of Donald Trump,” *Soc. Sci. Comput. Rev.*, vol. 36, no. 4, pp. 440–455, 2018, doi: 10.1177/0894439317730302.
[17] M. Conroy, J. T. Feezell, and M. Guerrero, “Facebook and political engagement: A study of online political group membership and offline political engagement,” *Comput. Human Behav.*, vol. 28, no. 5, pp. 1535–1546, Sep. 2012, doi: 10.1016/j.chb.2012.03.012.
[18] G. Hanson, P. M. Haridakis, A. W. Cunningham, R. Sharma, and J. D. Ponder, “The 2008 Presidential Campaign: Political Cynicism in the Age of Facebook, MySpace, and YouTube,” *Mass Commun. Soc.*, vol. 13, no. 5, pp. 584–607, Oct. 2010, doi: 10.1080/15205436.2010.513470.
[19] T. Macafee and J. J. De Simone, “Killing the Bill Online? Pathways to Young People’s Protest Engagement via Social Media,” *Cyberpsychology, Behav. Soc. Netw.*, vol. 15, no. 11, pp. 579–584, Nov. 2012, doi: 10.1089/cyber.2012.0153.
[20] K. D. Sweetser and R. W. Lariscy, “Candidates Make Good Friends: An Analysis of Candidates’ Uses of Facebook,” *Int. J. Strateg. Commun.*, vol. 2, no. 3, pp. 175–198, Jul. 2008, doi: 10.1080/15531180802178687.
[21] S. Stieglitz and L. Dang-Xuan, “Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior,” *J. Manag. Inf. Syst.*, vol. 29, no. 4, pp. 217–248, Apr. 2013, doi: 10.2753/MIS0742-1222290408.
[22] M. A. Xenos, T. Macafee, and A. Pole, “Understanding variations in user response to social media campaigns: A study of Facebook posts in the 2010 US elections,” *New Media Soc.*, vol. 19, no. 6, pp. 826–842, Jun. 2017, doi: 10.1177/1461444815616617.
[23] M. Bene, “Go viral on the Facebook! Interactions between candidates and followers on Facebook during the Hungarian general election campaign of 2014,” *Information, Commun. Soc.*, vol. 20, no. 4, pp. 513–529, Apr. 2017, doi: 10.1080/1369118X.2016.1198411.
[24] A. Rauchfleisch and J. Metag, “The special case of Switzerland: Swiss politicians on Twitter,” *New Media Soc.*, vol. 18, no. 10, pp. 2413–2431, Nov. 2016, doi: 10.1177/1461444815586982.
[25] R. Gerodimos and J. Justinussen, “Obama’s 2012 Facebook Campaign: Political Communication in the Age of the Like Button,” *J. Inf. Technol. Polit.*, vol. 12, no. 2, pp. 113–132, Apr. 2015, doi: 10.1080/19331681.2014.982266.
[26] A. O. Larsson, “Pandering, protesting, engaging. Norwegian party leaders on Facebook during the 2013 ‘Short campaign’’,’” *Information, Commun. Soc.*, vol. 18, no. 4, pp. 459–473, Apr. 2015, doi: 10.1080/1369118X.2014.967269.
[27] C. Bazerman and P. Prior, Eds., *What Writing Does and How It Does It*, 1st Editio. New York, USA: Routledge, 2003.
[28] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, *Probability & Statistics for Engineers & Scientists*, 8th ed. Pearson, 2006.
[29] CARVALHO, André Carlos Ponce de Leon Ferreira et al. Inteligência Artificial –Uma Abordagem de Aprendizado de Máquina – 2ª edição. Rio de Janeiro: LTC, 2021.3]
[30] HARRISON, Matt. Machine Learning Guia de Referência Rápida. São Paulo: Novatec, 2020
[31] Demertzis, Konstantinos. (2019). Classification VS Regression. 10.13140/RG.2.2.11389.51684.
[32] TAULLI, Tom. Introdução à Inteligência Artificial Uma abordagem não técnica. São Paulo: Novatec, 2020.
[33] Kumari, Khushbu & Yadav, Suniti. (2018). Linear regression analysis study. Journal of the Practice of Cardiovascular Sciences. 4. 33. 10.4103/jpcs.jpcs_8_18.
[34] Shanghai Arch Psychiatry. 2015; 27(2): 130-135. doi: http:/dx.doi.org/10.11919/j.issn.1002-0829.215044
[35] Cutler, Adele & Cutler, David & Stevens, John. (2011). Random Forests. 10.1007/978-1-4419-9326-7_5.
[36] Pamela J. Benoit & William L. Benoit (2005) Criteria for evaluating political campaign webpages, Southern Communication Journal, 70:3, 230-247, DOI: 10.1080/10417940509373329
[37] KLOSTERMAN, Stephen. Projetos de Ciência de Dados com Python. São Paulo: Novatec, 2019.
[38] Natekin, Alexey & Knoll, Alois. (2013). Gradient Boosting Machines, A Tutorial. Frontiers in neurorobotics. 7. 21. 10.3389/fnbot.2013.00021.
[39] Mishra, Sidharth & Sarkar, Uttam & Taraphder, Subhash & Datta, Sanjoy & Swain, Devi & Saikhom, Reshma & Panda, Sasmita & Laishram, Menalsh. (2017). Principal Component Analysis. International Journal of Livestock Research. 1. 10.5455/ijlr.20170415115235.