# Predicting Brazilian and U.S. Elections with Machine Learning and Social Media Data

Kellyton dos Santos Brito
*Centro de Informática[1], Departamento de Computação[2]*
*Universidade Federal de Pernambuco[1], Universidade Federal Rural de Pernambuco[2]*
Recife, Brazil
kellyton@kellyton.com.br

Paulo Jorge Leitão Adeodato
*Centro de Informática*
*Universidade Federal de Pernambuco*
Recife, Brazil
pjla@cin.ufpe.br

***Abstract*—Contemporary social networks, such as Facebook, Twitter, and Instagram, have reshaped the way politicians communicate with the electorate and run electoral campaigns. Scholars and the public have already perceived social media's strong impact on elections and its possible application to forecasting elections results. Current approaches focus on the volume of Twitter posts made by ordinary people talking about a candidate and use machine learning to identify the sentiment of these posts. However, differences regarding data collection, supporters' behavior on social networks, and the existence of bots can easily distort results. In this work, we propose a novel approach to training ML models for predicting vote share. This approach is based on modeling and using social media data gathered from the posts of official candidates' profiles combined with traditional polls. Then, we use an artificial neural network for prediction. Afterward, we perform experiments in two distinct scenarios: the 2018 Brazilian presidential election and the 2016 U.S. presidential election. The results show that the proposed approach has better vote share prediction results than polls in a scenario with many candidates and few available polls (Brazil) and is competitive in a scenario with two candidates and many available polls (U.S.). In addition, it shows advantages over many state-of-the-art approaches due to its simplicity, replicability, and robustness against volume manipulation. To the best of our knowledge, it is also the first attempt to validate ML models for the 2018 Brazilian election prediction.**

***Keywords*—Elections, Machine Learning, Neural Networks, Social Media, Social Networks, Facebook, Twitter, Instagram**

## I. Introduction

Social media (SM) have played a central role in politics and elections throughout this decade. We have entered a new era mediated by SM in which politicians conduct permanent campaigns without either geographic or time constraints, and extra information about them can be obtained not only by the press, but directly from their profiles on social networks (SNs) and through other people sharing and amplifying their voices on SM.

In this new scenario, SM is used extensively in campaigns, and an online campaign's success can even decide elections. Much academic research has been devoted to the modern political campaign and its activities [1], and how well Facebook and Twitter users reflect the general voting public [2]. In practice, recent examples of SN engagement and electoral success include the 2016 U.S. presidential election, when Donald Trump focused his campaign on free-media marketing [3], and the 2018 Brazilian presidential election, when the candidate with more SN engagement but little exposition on TV was elected [4].

Many initiatives focus on using SM data to predict elections outcomes [5]–[9]. However, predicting elections using SM has its challenges, such as a lack of historical data and barriers to data gathering. Following seminal approaches from Tumasjan [10] and O'Connor [11], most work is based on the volume of Twitter posts made by ordinary people talking about a candidate and uses machine learning (ML) to identify the sentiment of these posts. Then, researchers compare the volume of positive/negative tweets related to each candidate with election results or traditional polls. Despite alleged good results, these initiatives are constrained by the technical challenges to data gathering due to the high volume of Twitter posts, the under representativity of Twitter as SM, arbitrary choices of search keywords and timeframes, and the fact that differences regarding supporters' behaviors on SNs and the existence of bots can easily distort results, as criticized in the literature [12]–[14].

This paper proposes a novel approach to using ML models to predict elections outcomes. The approach is based on modeling and using SM data in a new form, focused on the repercussion of the posts of official candidates' profiles in all three major SNs (Facebook, Twitter, and Instagram). Then, repercussion data were combined with traditional polls and used to train ML models individually for each candidate to predict their vote share. For modeling, a multilayer perceptron (MLP) artificial neural network (ANN) was used, as well as traditional linear regression technique for baseline.

We conducted experiments in the scenarios of the 2018 Brazilian presidential election and the 2016 U.S. presidential election and evaluated the results statistically. Then, we compared our proposed approach with the last polls before the elections, as well as with the most recent state-of-the-art research. Our approach innovates by radically changing the SM input data modeling for prediction and by being independent of SNs. Furthermore, because of the reduced amount of data gathered for model training, its simplicity and replicability can be highlighted. It also innovates by training each candidate separately and being robust to varying supporter behaviors online and to the presence of bots. Finally, to the best of our knowledge, it is the first attempt to validate ML models to predict the 2018 Brazilian election.

## II. Related Research

Contemporary SN systems are new: Facebook launched to

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

the public in 2006, Twitter debuted in 2006, and Instagram emerged in 2010. However, the use of SM in modern political activities is already presenting promising results. In this section, we first explore SM's role in elections and the initial evidence of its correlation with electoral performance. Then, we briefly review the most-used models for predicting elections based on SM data, including their main challenges.

### A. The Use of Social Media in Elections and Electoral Performance

SM's impact on politics and elections is receiving ample attention from researchers worldwide. Smyth [15] studied how SM was used in the 2011 elections in West Africa, Nigeria, and Liberia, concluding that it helped to overcome scarcity of information during the electoral process. In the 2014 Indian general elections, Jaidka [16] identified new paradigms to engage and inform voters driven by modern information and communications technology (ICT). More recently, Aminolroya [17] highlighted that the flow of information from followees to followers on Instagram presented a significant role in the 2016 Iranian parliament election. Moreover, Morris' results [18] suggested that campaign messages about 2016 U.S. presidential candidates sent via Twitter—regardless of the candidate in focus—resonated just as strongly with potential voters as those sent via traditional media.

The correlation between SM performance and electoral results is also a focus of research. Kruikemeier's [19] results regarding Dutch national elections in 2010 showed that Twitter use had positive consequences for political candidates. In 2013, DiGrazia *et al.* [20] showed a statistically significant association between tweets that mentioned a candidate for the U.S. House of Representatives and their subsequent electoral performance. Later, Ramadhan [21] analyzed SM utilization in the 2014 Jakarta legislative election and showed that Facebook and Twitter usage was strongly correlated with the number of votes received by the candidates. More recently, Brito *et al.* [4] also found a strong correlation between user interaction on politicians' official SM profiles and their electoral performance during the 2018 Brazilian elections.

### B. Predicting Elections with Social Media Data

Because of the possible correlation between a politician's performance on SM and election outcomes, research on predicting election performance based on SM data has received attention in recent years.

Two studies can be considered seminal in this area. In 2010, Tumasjan *et al.* [10] presented a study on the 2009 German federal election. They collected all the tweets with the names of any of 6 parties represented in the German parliament or prominent politicians of these parties and compared the volume of tweets with the election results. As results, they claimed that "the mere number of tweets mentioning a political party can be considered a plausible reflection of the vote share and its predictive power even comes close to traditional election polls." In the same year and with an approach improved by sentiment detection of tweets, O'Connor [11] found that "a relatively simple sentiment detector based on Twitter data replicates consumer confidence and presidential job approval polls."

Based on these two studies, the volume of tweets combined with automatic sentiment detection became the main approach for most further research around the world [5]–[9]. In general terms, researchers collect posts on Twitter referring to a candidate or party; perform a sentiment analysis to classify the post as positive, negative, or neutral; and try to correlate the volume of positive and negative posts with electoral results. In these studies, the main challenges are gathering data via an open search on Twitter and the sentiment analysis. In fact, ML models are used in these studies mostly for sentiment analyses—not for prediction tasks.

These approaches engendered a number of criticisms [12]–[14] due to many limitations, such as (i) a lack of a replicable process; (ii) a lack of generalization because studies are performed only on one election; (iii) a lack of prediction capability during rallies; (iv) sample problems because Twitter cannot be generalized as a good sample of all SM and because gathered data do not represent even a good sample of all tweets; (v) arbitrary choices of data collection times and search keywords; and (vi) the possibility of being easily affected by volume manipulation from automated software, known as bots [22], spammers, paid propaganda, or even natural differences between users' behavior [23]. In fact, by using these approaches results can vary widely, as discussed by Jungherr [13], who after replicating Tumasjan's [10] seminal study, argued that "the results are contingent on arbitrary choices of the authors," and including just one more party or day of collection would greatly change the results.

## III. Methodology

This section presents the proposed methodology. Its scope is defined for presidential elections, focused on individuals rather than parties. We defined a process, based on CRISP-DM [24], that may be adapted for presidential elections all around the world. The process contains six phases: (i) business understanding, (ii) data understanding, (iii) data preparation, (iv) modeling, (v) evaluation, and (vi) deployment, the latter being a managerial phase. Due to the research nature, the deployment will not be addressed in this paper. All other phases are presented next.

### A. Business Understanding

Predicting elections with SM data has many differences and additional challenges compared to usual ML problems and solutions. First, the problem definition is not crystal clear from input data space to target definition. In some countries, such as the United States, a clear distinction exists among candidates of two traditional parties, and historical data can be gathered. However, in countries such as Brazil candidates and parties vary from one election to another, and almost no historical assumptions can be made. In such cases, each election should be analyzed independently.

In addition, almost no historical data about candidates exist and, worse, almost no labelled data are available for training. Only one actual labeled sample exists: the final vote share, which researchers want to predict before elections. Thus, one direction is to use traditional polls to train the models and the final vote share as the prediction. Due to this dependence, the selection of polls for training is also a challenge in two ways: different poll results according to pollsters and no evenly spaced time interval between polls, which usually decreases as election day approaches and imposes constraints in using traditional time-series approaches.

Moreover, the rapidly changing SM landscape must be considered. A given SN may be more prominent one year than in another year, and even in the same year one SN may be more relevant for one candidate but not for another.

Finally, concern about users' privacy has increased, especially after the Cambridge Analytica scandal [25]. Therefore, the use of users' personal data or profiling techniques, such as identifying individual users' vote preferences, should be avoided since access to those data is very likely to be prohibited in the future.

In this context, this proposal aims to predict candidates' final vote share in elections, which is a regression problem. The training and prediction will be based on SM data as features. Polling data will be used as labelled data for supervised training in time prior to elections. Then, the prediction will try to closely match raw election vote share, which has only one sample. Finally, performance will be measured as the mean and median of prediction errors compared with elections vote share.

### B. *Data Understanding*

In 1968, Zajonc [26] hypothesized that "mere repeated exposure of the individual to a stimulus object enhances his attitude toward it." Considering elections, in 1986 Oppenheimer made a correlation between politicians' exposition and electoral performance [27] and Mondak [28] also studied specific exposition in mass media. In the current study, we consider how the context of SM exposure affects elections.

Most studies consider how many people are talking about a candidate, but our hypothesis considers how many people are paying attention to a candidate and propagating his or her presence. In propagating their presence, we consider two characteristics of SM: (i) a user can directly share politicians' content, republishing it for friends or followers, and (ii) SN algorithms prioritize showing content with more engagement, creating a snowball effect [29]. Thus, we consider as the main data input for this study people's engagement on candidates' official profiles, as follows: (i) Facebook: number of likes, shares, and comments on candidates' posts; (ii) Twitter: number of likes and retweets of candidates' tweets; and (iii) Instagram: number of likes and comments on candidates' posts. If another relevant SN is identified, then it can also be added by following the same rationale of interactions in future elections. For example, considering YouTube, the number of visualizations, likes, and comments would be considered.

Understanding polling data is a higher challenge. Some countries, such as United States, have a high number of publicly available polls and daily weighted averages created by news companies, such as the ones created by Huffington Post (HP) [30] and New York Times (NYT) [31]. Nevertheless, in many countries the access to polls is a barrier. In Brazil, few polls are made publicly available, in Mexico data must be manually gathered from the national repository, and in Uruguay, data must be collected directly from newspaper websites.

Finally, considering that poll results vary by methodology and pollster, a strategy for using these data must be defined. Possible options include the use of a weighted average calculated by another institution (*e.g.*, HP, NYT), the calculation of a new weighted average, the selection of most trusted pollsters, the exclusion of polls with outlier results or other criteria.

### C. *Data Preparation*

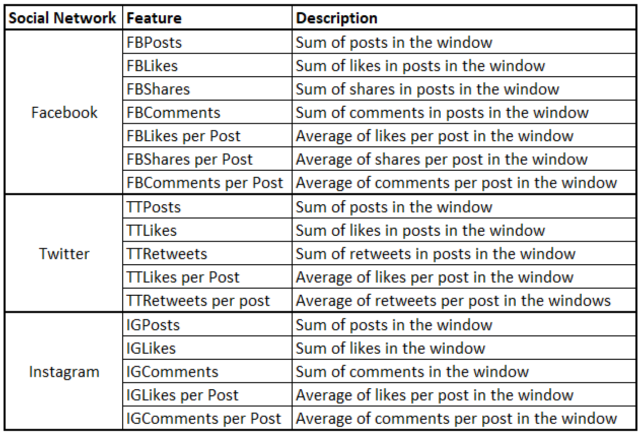
Because SM data used in this approach is public data collected from politicians' official profiles, it can be collected using official application programming interfaces (APIs). As a result, the data are complete and data cleaning is not necessary. However, the initial dataset is enhanced and completely transformed to be used in the proposed ML approach.

Data are modeled so that the result ***r*** of a poll (or the election results) at a specific date ***d*** is a function of the engagement observed in the candidate's SNs—Facebook (***F***), Twitter (***T***), and Instagram (***I***)—in an aggregate window of ***w*** days prior ***d***, as presented in (1).

$$r(d_w) = f(F, T, I)_{d\text{-}w..d\text{-}1} \qquad (1)$$

This initial dataset is also enhanced by the addition of new variables to the original, by not only counting the total number of interactions, but correlating the number of interactions per post. Then, we reach a final dataset with 17 features, as presented in Table I.

TABLE I. FINAL LIST OF FEATURES



| Social Network | Feature | Description |
|---|---|---|
| Facebook | FBPosts | Sum of posts in the window |
| | FBLikes | Sum of likes in posts in the window |
| | FBShares | Sum of shares in posts in the window |
| | FBComments | Sum of comments in posts in the window |
| | FBLikes per Post | Average of likes per post in the window |
| | FBShares per Post | Average of shares per post in the window |
| | FBComments per Post | Average of comments per post in the window |
| Twitter | TTPosts | Sum of posts in the window |
| | TTLikes | Sum of likes in posts in the window |
| | TTRetweets | Sum of retweets in posts in the window |
| | TTLikes per Post | Average of likes per post in the window |
| | TTRetweets per post | Average of retweets per post in the windows |
| Instagram | IGPosts | Sum of posts in the window |
| | IGLikes | Sum of likes in the window |
| | IGComments | Sum of comments in the window |
| | IGLikes per Post | Average of likes per post in the window |
| | IGComments per Post | Average of comments per post in the window |

As an example, for a poll published on January 30 with a window of 28 days, input data are the individualized sum of all the posts, likes, comments, and shares/retweets from Facebook, Twitter, and Instagram from January 2 to January 29, and the ratio "per post" per SN for all of them.

Finally, considering the small number of polls, resulting in a small number of data samples for training, it is desirable to use feature selection or dimensionality reduction techniques, such as principal component analysis (PCA), to avoid well-known problems of high dimensionality [32] and violation of VC-dimension [33], which are present in several papers referenced above. PCA is desirable because it eliminates the collinearity among features, which is likely in this scenario, while allowing dimensionality reduction.

### D. *Modeling*

The candidate vote share prediction problem has been characterized as a regression problem. Due to the existence of many regression techniques, we selected traditional linear regression as baseline, as well as a more sophisticated and nonlinear technique, multi-layer perceptron (MLP) artificial neural network (ANN). For context, Zolghadr *et al.* [34] presented four main points that justify the use of ANNs to forecast presidential elections: (i) ANNs can capture nonlinear relations between independent (input) and dependent (output) variables; (ii) ANNs are data driven, so no explicit assumption is needed for the model between the inputs and outputs; (iii) ANNs can generalize, producing good results even when they face new input patterns; and (iv) ANNs do not need assumptions on the distribution of input data, unlike statistical techniques. Finally, MLP can solve complex problems stochastically and is a universal function approximator [35].

One challenge of using MLP is tuning its parameters. In this sense, it is desirable to choose parameters by selecting similar problems in the literature or using techniques for automatic selection of parameters such as grid or random parameter search.

### E. *Evaluation*

Evaluation must measure the difference between predicted results, based on training with SM and polling data, and each candidate's final vote share. For this, we consider two metrics: mean absolute error (MAE), which measures the absolute error, and the mean absolute percentage error (MAPE), which measures the percentage error. Although most related studies use only MAE, we consider MAPE relevant because, for example, an error of 3 points in vote share is much more relevant for a candidate with 2% of votes than for a candidate with 50% of votes, and this relevance is not captured by MAE.

For comparison, MAE and MAPE of predictions using our approach is compared with MAE and MAPE of predictions of the last published polls before election's day. In addition, the paired Wilcoxon's signed-rank test is applied on errors to verify if statistically significant differences exist between results.

## IV. Experiments

For evaluation of the proposed approach, we performed two experiments. The first used 2018 Brazilian presidential first round elections data and the second used data from the 2016 U.S. presidential election. Although vote share is not the main point of U.S. elections, this second experiment does not aim to predict the elected candidate, but it allows comparison and validation of results of the first experiment, as well as the comparison with other studies found in the literature.

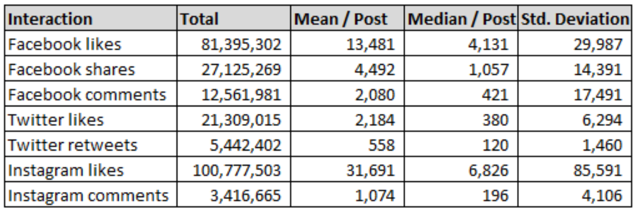
### A. *Predicting Brazilian Elections*

*1) Business Understanding:* The first round of Brazilian presidential elections was held on October 7, 2018. There were 13 candidates, five of them being considered the main candidates who received more than 1% of votes. This experiment will consider only these five candidates. A recent study of this election [4] showed a correlation between engagement and electoral performance, but it was not able to model this correlation. In fact, the candidate with the second most interactions on SN received the fifth most votes.

Regarding polls, 14 institutes are listed as pollsters, but only two of them are well-known as the most trustworthy, Ibope and Datafolha, whose polls were used in this experiment.

*2) Data Understanding:* The five candidates made 18,976 posts on SM from January 1 until election day. Most posts were on Twitter (51%), followed by Facebook (32%) and Instagram (17%). However, Twitter posts received the fewest number of interactions. Candidates' posts generated 252 million interactions as shown in Table II, which also shows the difference in interactions performed in each SN, the mean and median, and the high standard deviation of each type of interaction.

TABLE II. Interactions with Brazilian Candidates' Posts on Social Networks



| Interaction | Total | Mean / Post | Median / Post | Std. Deviation |
|---|---|---|---|---|
| Facebook likes | 81,395,302 | 13,481 | 4,131 | 29,987 |
| Facebook shares | 27,125,269 | 4,492 | 1,057 | 14,391 |
| Facebook comments | 12,561,981 | 2,080 | 421 | 17,491 |
| Twitter likes | 21,309,015 | 2,184 | 380 | 6,294 |
| Twitter retweets | 5,442,402 | 558 | 120 | 1,460 |
| Instagram likes | 100,777,503 | 31,691 | 6,826 | 85,591 |
| Instagram comments | 3,416,665 | 1,074 | 196 | 4,106 |

Collected polling data were published by Ibope and Datafolha from January 1, 2018, until the day before the election. There are in total 21 polls, 11 from Datafolha starting from January 30, and 10 from Ibope starting from July 24.

*3) Data Preparation:* Because the final list of candidates was only known at the official launch of candidatures (August 5), polls performed before this date contained many possible scenarios, and scenarios closest to the final candidate list were chosen.

A set of 10 independent datasets was generated for each of the five candidates considered, with the features explained in Section 3.c. Each dataset was generated with a different aggregated window, w = [1..7, 14, 21, 28], to avoid an arbitrary window selection, which would introduce bias into the experiments.

Due to the high number of features for the few training samples, PCA was applied on each dataset independently, using scikit-learn libraries. Component selection was set to cover a variance higher than 95%. Thus, the number of components varied from four to seven in a 1-day window, and from two to three in a 28-day window, drastically reducing the number of input features to prevent the well-known problem of high dimensionality.

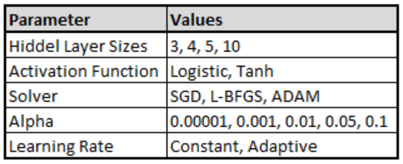
*4) Modeling:* The new datasets were applied to a traditional MLP-BP artificial network. It was implemented using the Python scikit-learn libraries. The ANN was trained with data until the last prediction 1 day before elections (data from 21 polls). Then, it made one prediction for the final vote share, and results were compared with actual vote share. For parameter setting, two approaches were used: manual parameter selection and grid search for parameters. By considering data characteristics, mainly small sample, we manually chose the following parameters: one hidden layer with three neurons, to avoid overfitting; L-FBGS as solver, which has good performance with small sample; alpha set to

0.05 and constant learning rate for fast training, and logistic activation. For grid search, we used parameters presented in Table III. One disadvantage of this approach is that a different result may be achieved each time this sequence is performed. To avoid this execution bias, each execution was repeated five times, and we used the mean value as the actual predicted vote share.

Ten predictions were run with each parameter selection method (fixed parameters and grid search), one for each aggregated window size, w = [1..7, 14, 21, 28]. It is a well-known principle that averaging the output of several networks may give us a better and more stable result [36]. Thus, to reduce variance we used a combined approach, by using the mean and median of all 10 predictions as the final prediction for each candidate, in a simplified implementation of a committee machine [37].

As baseline technique, a linear regression was also applied in the same datasets. Fig. 1 illustrates data preparation and modelling main steps.

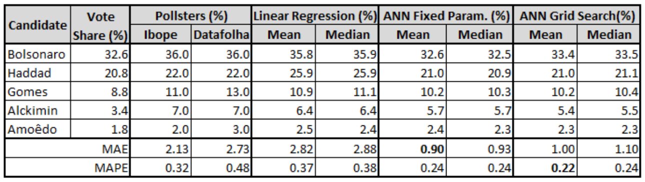
TABLE III. VALUES FOR ANN GRID SEARCH PARAMETERS



| Parameter | Values |
|---|---|
| Hiddel Layer Sizes | 3, 4, 5, 10 |
| Activation Function | Logistic, Tanh |
| Solver | SGD, L-BFGS, ADAM |
| Alpha | 0.00001, 0.001, 0.01, 0.05, 0.1 |
| Learning Rate | Constant, Adaptive |

5) *Evaluation:* Evaluation was performed by comparing MAE and MAPE metrics of predicted vote share with the most recent polling results. Results using MLP-BP with fixed parameters and with grid search parameters were evaluated, as well as results with linear regression (LR). Prediction results and errors are shown in Table IV, including the most recent polling data before elections and official raw vote shares.

TABLE IV. PREDICTIONS AND ERRORS FOR EACH BRAZILIAN CANDIDATE, COMPARED WITH FINAL VOTE SHARE



| Candidate | Vote Share (%) | Pollsters (%) | | Linear Regression (%) | | ANN Fixed Param. (%) | | ANN Grid Search(%) | |
|---|---|---|---|---|---|---|---|---|---|
| | | Ibope | Datafolha | Mean | Median | Mean | Median | Mean | Median |
| Bolsonaro | 32.6 | 36.0 | 36.0 | 35.8 | 35.9 | 32.6 | 32.5 | 33.4 | 33.5 |
| Haddad | 20.8 | 22.0 | 22.0 | 25.9 | 25.9 | 21.0 | 20.9 | 21.0 | 21.1 |
| Gomes | 8.8 | 11.0 | 13.0 | 10.9 | 11.1 | 10.2 | 10.3 | 10.2 | 10.4 |
| Alckimin | 3.4 | 7.0 | 7.0 | 6.4 | 6.4 | 5.7 | 5.7 | 5.4 | 5.5 |
| Amoêdo | 1.8 | 2.0 | 3.0 | 2.5 | 2.4 | 2.4 | 2.3 | 2.3 | 2.3 |
| | MAE | 2.13 | 2.73 | 2.82 | 2.88 | **0.90** | 0.93 | 1.00 | 1.10 |
| | MAPE | 0.32 | 0.48 | 0.37 | 0.38 | 0.24 | 0.24 | **0.22** | 0.24 |

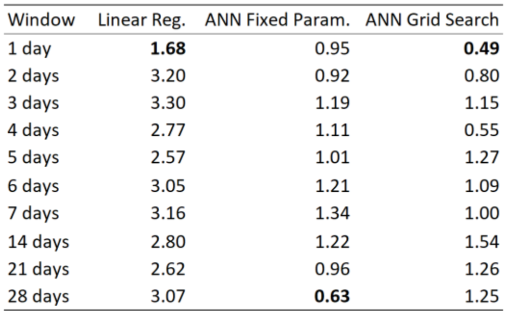
Best results were achieved with ANN: best MAE with fixed parameters and best MAPE with grid searched parameters. In fact, whether or not grid search was used, using the mean or median achieved very close results. Moreover, all these results were better than the most recent polls of traditional pollsters. In addition, the worst results considering MAE were obtained with LR, and considering MAPE LR results were only better than Datafolha. Finally, error results using mean or median of aggregated windows were almost the same.

In addition to MAE and MAPE, the statistical significance Wilcoxon's signed-rank test was applied on absolute error (AE) of each candidate. The test was applied by comparing each method (ANN with fixed parameters, with grid search, and LR) with each other and with the most recent polls, with the three alternative hypotheses: equals results, higher than results, and lower than results.

Results present statistical support ($p < 0.05$) to claim that errors obtained with both ANN strategies were lower than errors of Datafolha polls and the LR method ($p = 0.031$ on all four tests). In addition, when similarity of errors obtained by ANN with fixed and grid search parameters was tested, the p-value was $p = 1$, suggesting that values are almost identical. On the other hand, there is no statistical evidence to claim that errors obtained with either of the two ANN approaches are lower than errors of Ibope polls, despite MAE and MAPE being lower.

A posterior analysis shows that a prediction using 1-day window could have had the best result in the ANN with grid search parameters setup, with a MAE of 0.49, and a 28-day window could have had the best result in ANN with Fixed parameters setup (MAE of 0.63). However, we do not claim this result as the result of our study because it could not be achieved in a reliable way before knowing election's results. In fact, due to the high variation, as shown in Table V, our data does not allow us to make generalizable conclusions about the best window size, and we argue that the committee strategy is preferable.

TABLE V. MAE ERRORS OBTAINED WITH DIFFERENT WINDOWS



| Window | Linear Reg. | ANN Fixed Param. | ANN Grid Search |
|---|---|---|---|
| 1 day | **1.68** | 0.95 | **0.49** |
| 2 days | 3.20 | 0.92 | 0.80 |
| 3 days | 3.30 | 1.19 | 1.15 |
| 4 days | 2.77 | 1.11 | 0.55 |
| 5 days | 2.57 | 1.01 | 1.27 |
| 6 days | 3.05 | 1.21 | 1.09 |
| 7 days | 3.16 | 1.34 | 1.00 |
| 14 days | 2.80 | 1.22 | 1.54 |
| 21 days | 2.62 | 0.96 | 1.26 |
| 28 days | 3.07 | **0.63** | 1.25 |

In a similar way, it was not possible to make conclusive assumptions regarding the number of networks used on the input model and results. Table VI shows errors using different combinations of SN in a window of 28 days, lower MAE errors may be achieved either by using data from only one SN, or when using the combination of all three. These data suggest that it would not be feasible to define *a priori* which combination of social networks is the most suitable to use for prediction. As a consequence, the strategy of letting the neural network identify the weights automatically seems to be more appropriate.

### B. *Predicting U.S. Elections*

1) *Business Understanding:* The U.S. elections were held on November 8, 2016. Historically, only two candidates are considered the main candidates, running for the Democratic Party and the Republican Party, and these two were considered in this experiment. Also, the U.S. presidential election is indirect, and a president can be elected despite not having the majority of the popular vote. Therefore, our proposed approach does not aim to predict the elected president, but the final popular vote share of each candidate.

Regarding polls, there is a great availability of polls from many pollsters that present very different results. The polling data are also publicly available (*e.g.*, the Huffington Post Pollster API) [38]. Using these data, other institutions publish daily weighted averages, such as the New York Times [31]. In this study, we used the daily average published by the Huffington Post [30], since it is the main polling data publisher.

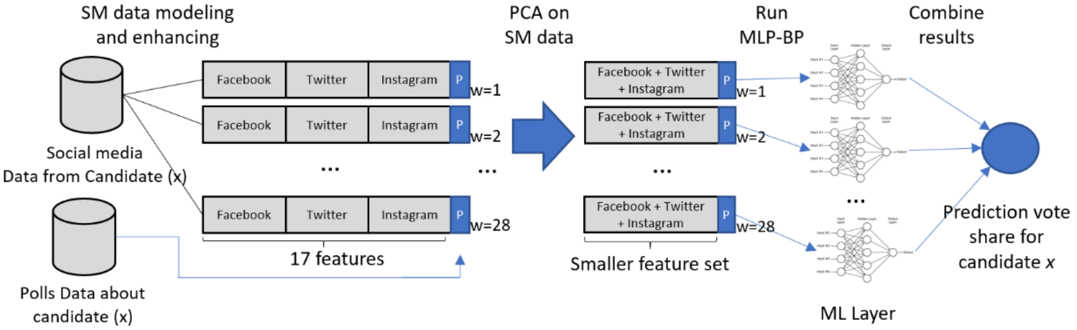


Fig. 1. Main steps of data preparation and modelling

Due to the availability of polling data, we used SM and polling data starting from 1 year before elections, November 08, 2015, until November 07, 2016.

TABLE VI. MAE ERRORS OBTAINED WITH DIFFERENT COMBINATIONS OF SOCIAL NETWORKS IN A WINDOW OF 28 DAYS

| Model | Linear Reg. | ANN Fixed Param. | ANN Grid Search |
|---|---|---|---|
| Facebook | 3.37 | 1.16 | 0.89 |
| Twitter | **2.60** | 0.87 | 0.85 |
| Instagram | 3.34 | 0.90 | **0.79** |
| Facebook + Twitter | 3.10 | 1.19 | 0.94 |
| Facebook + Instagram | 3.35 | 1.26 | 1.22 |
| Twitter + Instagram | 2.93 | 1.05 | 1.15 |
| All three SN | 3.07 | **0.63** | 1.25 |

*2) Data Understanding:* In the studied period, the two candidates made 12,558 SM posts. Most posts were issued on Twitter (76.3%), followed by Instagram (12.0%) and Facebook (11.7%). However, as occurred in Brazil, Twitter posts received fewer interactions for each post. In total, candidates' posts generated 349 million interactions as shown in Table VII, which also shows the difference in interactions performed in each SN, the mean and median, and the high standard deviation of each type of interaction. Regarding polls, as we used the daily average published by the Huffington Post in the period of 1 year, 366 polls (2016 was a leap year) were used.

*3) Data Preparation and Modeling*: Data preparation and modeling was performed the same as in experiment 1.

*4) Evaluation:* Evaluation was performed the same as in experiment 1. Prediction results are shown in Table VIII.

MAE and MAPE results surprisingly show lower errors obtained by LR, with better results than the last Huffington prediction. Using ANN with grid search for parameters, errors were close to Huffington Post's last poll, and ANN with fixed parameters obtained higher MAE and MAPE errors. As occurred in the Brazilian experiment, results using mean or median of aggregated windows were almost the same.

Considering the statistical test, it is more difficult to achieve statistical support because there are only two samples (Clinton and Trump absolute errors). Indeed, the same statistical tests were applied as in Brazilian experiment, but no statistical conclusions can be made because no tests presented a p-value lower than 0.05. These data suggest that our prediction errors are similar to errors obtained by Huffington Post's last poll.

TABLE VII. INTERACTIONS WITH U.S. CANDIDATES' POSTS ON SOCIAL NETWORKS

| Interaction | Total | Mean / Post | Median / Post | Std. Deviation |
|---|---|---|---|---|
| Facebook likes | 82,973,796 | 56,560 | 38,029 | 59,542 |
| Facebook shares | 18,796,209 | 12,813 | 5,050 | 34,411 |
| Facebook comments | 21,663,422 | 14,767 | 5,512 | 49,221 |
| Twitter likes | 97,914,549 | 10,223 | 5,553 | 14,820 |
| Twitter retweets | 38,541,629 | 4,024 | 2,171 | 8,028 |
| Instagram likes | 79,789,762 | 52,736 | 42,859 | 40,274 |
| Instagram comments | 9,125,977 | 6,032 | 4,155 | 8,238 |

TABLE VIII. PREDICTIONS AND ERRORS FOR EACH U.S. CANDIDATE, COMPARED WITH FINAL VOTE SHARE

| Candidate | Vote | Polls | Linear Regression (%) | | ANN Fixed Param. (%) | | ANN Grid Search (%) | |
|---|---|---|---|---|---|---|---|---|
| | Share (%) | Huffpost | Mean | Median | Mean | Median | Mean | Median |
| Trump | 46.1 | 42.0 | 42.6 | 42.5 | 40.2 | 40.4 | 41.6 | 41.6 |
| Clinton | 48.2 | 47.3 | 48.2 | 48.3 | 46.9 | 46.7 | 47.2 | 47.6 |
| | MAE | 2.50 | **1.78** | 1.85 | 3.58 | 3.58 | 2.80 | 2.56 |
| | MAPE | 0.05 | **0.04** | 0.04 | 0.08 | 0.08 | 0.06 | 0.06 |

## V. DISCUSSION AND COMPARISON OF RESULTS

The experiments show that combining SM data with traditional polls may have similar, or even better, results than traditional polls alone. Results were better in the Brazilian scenario, with few polls and more candidates, than in the U.S. scenario, with more available polls and only two main candidates. By having numerous available polls from many pollsters and a mix of strategies (automated or live phone calls, online polls, and in-person polls), daily weighted averages can capture both the opinions of people who support politicians on SM as well as of people who do not show their opinions online. Then, the introduction of SM-gathered data had a small effect on U.S. election prediction.

On the other hand, in the Brazilian scenario where polls are scarcer and are performed mainly in-person, the addition of SM data helps to improve poll results. In addition, due to the sparse and variable time elapsed between polls, this approach can be useful to estimate vote share in the interval between them. For example, in the campaign period, models can be trained with available past polling data and used to estimate the vote share on a daily basis.

As presented in Section II, many methods consist of counting the number of users mentioning each candidate per day. These methods present many challenges, since the choice of keywords related to each candidate and the choice of data gathering period can interfere with the results. In addition, the behavioral differences of each candidate's supporters or the presence of bots may also affect results. Finally, these methods are based solely on Twitter. In the proposed approach, these challenges are addressed: data are collected from candidates' official profiles, thus no keyword selection is needed; we collect data from at least 6 months before elections and use many time windows for data analysis, ranging from 1 to 28 days before target date, using a combined result; and the individual training of each candidate model with polls helps reduce the influence of bias in supporters' behavior or the existence of bots. In addition, all three major SNs are considered, and the addition or removal of any other SN is trivially possible.

Another widely used method consists of adding sentiment analysis to the mentioned counts. In addition to the aforementioned challenges, it introduces the selection and training of an appropriate method for sentiment analysis. This challenge is softened in our approach since interactions on a candidate's official profile usually demonstrate support. Moreover, even in cases where comments would mean disagreement, the training and weights calculation would capture this behavior and set appropriate positive or negative weights to each feature.

Moreover, the presented approach collects much less data, thousands of posts from less than a dozen of candidates, instead of millions of posts from the entire population that are collected by other methods. This characteristic is becoming more important due to increased limitations on data gathering in SNs. In addition, this limited and direct data gathering improves the approach's replicability.

We consider two main challenges in this study that may be addressed in future research. First is the poll selection, which may directly affect the results. Approaches for pruning polls, such as discarding outliers, would be desirable. Second, the ML algorithm and parameter selection may also be improved, such as the use of other ML approaches focused on small samples.

We may compare our approach and results for the 2016 U.S. election with the study presented in [7], where researchers collected 171 million tweets and experimented with four methods for generating opinion time series. As result, they claimed one of their four methods produced a good fit with the polls by a reduced root-mean-square-error, but that was not statistically validated. The main difference of this study and our approach is the amount of data necessary to obtain similar results: 171 million tweets versus 12,000 posts. Still regarding the 2016 U.S. elections, Heredia *et al.* [6] collected 3 million tweets and used both volume and sentiment for prediction election results, and they compared results with three pollsters. Their findings show that neither volume nor sentiment are predictors of election outcomes or polling at the state level. At the national level, considering one specific time window (65 days before elections), their method matched one of the three selected pollsters. However, if considering any other time window or mean or median of predictions, their results are far from both polls and final vote share, performing worse than our approach.

Considering studies analyzing multiple elections, Anjaria *et al.* [8] employed four ML techniques for sentiment analysis and combined results with influence factor based on retweets, to predict 2012 U.S. presidential elections and 2013 Karnataka (India) state assembly elections. Regarding U.S. results, their best result presented a MAE of 3.44, which is comparable with our worst result. Nevertheless, they obtained a MAE of 13.60 on Indian elections considering four parties, which largely differs from actual vote share.

Finally, Tsakalidis *et al.* [5] used Twitter data to predict the 2014 EU election results in Germany, the Netherlands, and Greece. We consider this work the closest to ours. Despite basing their study on Twitter volume and sentiment, they used 11 derived variables combined with one poll-based feature. Three algorithms were applied for regression, linear regression, Gaussian process, and sequential minimal optimization, and the output average was used as the final estimate in this combined regressor. The authors used 26 polls from Greece, nine from Germany, and 13 from the Netherlands. Based on MAE and MSE, their approach performed best on German and Greek elections, but results for the Netherlands was worse than the average computed by the website where they collected polling data. Good results in two elections are in line with our argument that generating domain-based derived variables and training ML algorithms with these data combined with polls can achieve good results. One of the main challenges of the study was the selection of the training window, they used an arbitrary 7-day window, despite having tested other values. Finally, their approach also needed many more posts than ours (361,713, 452,348, and 263,465 tweets for Germany, the Netherlands, and Greece, respectively).

## VI. Concluding Remarks and Future Works

This paper proposed an approach to training ML models for predicting vote share. This approach is based on modeling and using SM data gathered from the posts on candidates' official profiles combined with traditional polls. Data from candidates' posts were transformed and enhanced by the creation of aggregation windows and of new variables, such as number of likes, shares, and comments per post. Then, we trained ML regression models, linear regression, and MLP-BP ANN, and predicted results of presidential elections in Brazil (2018) and the United States (2016).

Based on MAE and MAPE error metrics, the proposed approach has better vote share prediction results than polls in a scenario with many candidates and few available polls (Brazil) and is competitive in a scenario with two candidates and many available polls (United States). In addition, statistical tests show that Brazilian predictions were better than those of one of the pollsters used for training, Datafolha, and similar to another pollster, Ibope. Moreover, U.S. predictions show results close to the weighted averages published by the Huffington Post.

The presented approach innovates the source data used for model training, based on candidates' official profiles, and collects much fewer posts than state-of-the-art approaches. It also uses data from all major SNs, instead of being Twitter-focused, and new SNs can be easily included or excluded. It also innovates by training each candidate separately and being robust to varying supporter behaviors online and the presence of bots. Moreover, the common bias regarding arbitrary selection of keywords and time interval for search is avoided. Finally, to the best of our knowledge, it is the first attempt to validate ML models for 2018 Brazilian election prediction.

Challenges for future work were also identified. The first challenge is the selection of input polls, which may affect results. Also, the improvement of ML algorithms, such as the use other ML approaches focused on small training sets, would be desirable.

## REFERENCES

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

[2] J. Mellon and C. Prosser, "Twitter and Facebook are not representative of the general population: Political attitudes and demographics of british social media users," Res. Polit., vol. 4, no. 3, 2017.

[3] 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.

[4] K. Brito, N. Paula, M. Fernandes, and S. Meira, "Social Media and Presidential Campaigns – Preliminary Results of the 2018 Brazilian Presidential Election," in 20th Annual International Conference on Digital Government Research on - dg.o 2019, 2019, pp. 332–341.

[5] A. Tsakalidis, S. Papadopoulos, A. I. Cristea, and Y. Kompatsiaris, "Predicting Elections for Multiple Countries Using Twitter and Polls," IEEE Intell. Syst., 2015.

[6] B. Heredia, J. D. Prusa, and T. M. Khoshgoftaar, "Social media for polling and predicting United States election outcome," Soc. Netw. Anal. Min., vol. 8, no. 1, p. 48, Dec. 2018.

[7] A. Bovet, F. Morone, and H. A. Makse, "Validation of Twitter opinion trends with national polling aggregates: Hillary Clinton vs Donald Trump," Sci. Rep., 2018.

[8] M. Anjaria and R. M. R. Guddeti, "A novel sentiment analysis of social networks using supervised learning," Soc. Netw. Anal. Min., vol. 4, no. 1, p. 181, Dec. 2014.

[9] N. D. Prasetyo and C. Hauff, "Twitter-based Election Prediction in the Developing World," Proc. 26th ACM Conf. Hypertext Soc. Media - HT '15, 2015.

[10] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," in Fourth International AAAI Conference on Weblogs and Social Media, 2010.

[11] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in 4th International AAAI Conference on Weblogs and Social Media, 2010.

[12] D. Gayo-Avello, P. T. Metaxas, and E. Mustafaraj, "Limits of Electoral Predictions using Social Media Data," Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media. 2011.

[13] A. Jungherr, P. Jürgens, and H. Schoen, "Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, A., sprenger, T. O., sander, P. G., & welpe, I. M. 'predicting elections with twitter: What 140 characters reveal about political sentiment,'" Soc. Sci. Comput. Rev., 2012.

[14] A. Jungherr, H. Schoen, O. Posegga, and P. Jürgens, "Digital Trace Data in the Study of Public Opinion," Soc. Sci. Comput. Rev., vol. 35, no. 3, pp. 336–356, Jun. 2017.

[15] J. DiGrazia, K. McKelvey, J. Bollen, and F. Rojas, "More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior," PLoS One, vol. 8, no. 11, p. e79449, Nov. 2013.

[16] K. Jaidka and S. Ahmed, "The 2014 Indian general election on Twitter: an analysis of changing political traditions," in Proceedings of the Seventh International Conference on Information and Communication Technologies and Development, 2015.

[17] Z. Aminolroaya and A. Katanforoush, "How Iranian Instagram users act for parliament election campaign A study based on followee network," in 2017 3rd International Conference on Web Research, ICWR 2017, 2017.

[18] D. S. Morris, "Twitter Versus the Traditional Media: A Survey Experiment Comparing Public Perceptions of Campaign Messages in the 2016 U.S. Presidential Election," Soc. Sci. Comput. Rev., vol. 36, no. 4, pp. 456–468, 2018.

[19] S. Kruikemeier, "How political candidates use Twitter and the impact on votes," Comput. Human Behav., vol. 34, pp. 131–139, 2014.

[20] J. DiGrazia, K. McKelvey, J. Bollen, and F. Rojas, "More tweets, more votes: Social media as a quantitative indicator of political behavior," PLoS One, 2013.

[21] D. A. Ramadhan, Y. Nurhadryani, and I. Hermadi, "Campaign 2.0: Analysis of social media utilization in 2014 Jakarta legislative election," in Proceedings - ICACSIS 2014: 2014 International Conference on Advanced Computer Science and Information Systems, 2014.

[22] A. Bessi and E. Ferrara, "Social bots distort the 2016 U.S. Presidential election online discussion," First Monday, vol. 21, no. 11, Nov. 2016.

[23] E. Mustafaraj, S. Finn, C. Whitlock, and P. T. Metaxas, "Vocal minority versus silent majority: Discovering the opionions of the long tail," in Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011, 2011.

[24] C. Shearer, "The CRISP-DM model: the new blueprint for data mining," J. data Warehous., vol. 5, no. 4, pp. 13--22, 2000.

[25] J. Isaak and M. J. Hanna, "User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection," Computer (Long. Beach. Calif)., vol. 51, no. 8, pp. 56–59, Aug. 2018.

[26] R. B. ZAJONC, "Attitudinal Effects Of Mere Exposure," J. Pers. Soc. Psychol., vol. 9, no. 2, 1968.

[27] B. I. Oppenheimer, J. A. Stimson, and R. W. Waterman, "Interpreting U. S. Congressional Elections: The Exposure Thesis," Legis. Stud. Q., vol. 11, no. 2, pp. 227–247, 1986.

[28] J. J. Mondak, "Media exposure and political discussion in U.S. elections," J. Polit., vol. 57, no. 1, 1995.

[29] The Facebook, "News Feed FYI: A Window Into News Feed," 2013. [Online]. Available: https://www.facebook.com/business/news/News-Feed-FYI-A-Window-Into-News-Feed. [Accessed: 01-Nov-2019].

[30] The Huffington Post, "2016 General Election: Trump vs. Clinton," 2016. [Online]. Available: https://elections.huffingtonpost.com/pollster/2016-general-election-trump-vs-clinton. [Accessed: 01-Nov-2019].

[31] The New York Times, "The NYT Last Elections Polls 2016," 2016. [Online]. Available: https://www.nytimes.com/interactive/2016/us/elections/polls.html. [Accessed: 01-Nov-2019].

[32] G. V. Trunk, "A Problem of Dimensionality: A Simple Example," IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-1, no. 3, pp. 306–307, Jul. 1979.

[33] V. Vapnik, E. Levin, and Y. Le Cun, "Measuring the VC-Dimension of a Learning Machine," Neural Comput., vol. 6, no. 5, pp. 851–876, Sep. 1994.

[34] M. Zolghadr, S. A. A. Niaki, and S. T. A. Niaki, "Modeling and forecasting US presidential election using learning algorithms," J. Ind. Eng. Int., 2018.

[35] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," Neural Networks, vol. 2, no. 5, pp. 359–366, Jan. 1989.

[36] L. Breiman, "Bagging predictors," Mach. Learn., vol. 24, no. 2, pp. 123–140, Aug. 1996.

[37] V. Tresp, "Committee Machines," in Handbook of Neural Network Signal Processing, Y. H. Hu and J.-N. Hwang, Eds. CRC Press, 2001.

[38] The Huffington Post, "Pollster API," 2016. [Online]. Available: http://elections.huffingtonpost.com/pollster/api/v2. [Accessed: 01-Nov-2019].