RESEARCH

Open Access

Inferring the votes in a new political landscape: the case of the 2019 Spanish Presidential elections

Didier Grimaldi[1]* , Javier Diaz Cely[2] and Hugo Arboleda[2]

*Correspondence:
didier.grimaldi@salle.url.edu
[1] Department of Strategy,
Entrepreneurship
and Innovation, La Salle
Campus, Ramon Llull
University, Sant Joan de La
Salle 42, 08022 Barcelona,
Spain
Full list of author information
is available at the end of the
article

Abstract

The avalanche of personal and social data circulating in Online Social Networks over the past 10 years has attracted a great deal of interest from Scholars and Practitioners who seek to analyse not only their value, but also their limits. Predicting election results using Twitter data is an example of how data can directly influence the politic domain and it also serves an appealing research topic. This article aims to predict the results of the 2019 Spanish Presidential election and the voting share of each candidate, using Tweeter. The method combines sentiment analysis and volume information and compares the performance of five Machine learning algorithms. Several data scrutiny uncertainties arose that hindered the prediction of the outcome. Consequently, the method develops a political lexicon-based framework to measure the sentiments of online users. Indeed, an accurate understanding of the contextual content of the tweets posted was vital in this work. Our results correctly ranked the candidates and determined the winner by means of a better prediction of votes than official research institutes.

Keywords: Machine learning, Election, Prediction, Spain, Twitter, Big social data, Online Social Networks

Introduction

For more than a decade now, with the emergence of Internet 2.0, users have been able to generate their own content and share it publicly with relative ease. This boom has witnessed a huge surge in the popularity of Online Social Networks (OSN), in particular the Twitter microblogging platform which allows its users to share text messages of up to 140 characters with their family, friends and followers [12]. More than 500 million messages, commonly known as tweets, are posted every day [20, 34]. Given that the main reason behind these publications is to express the opinion of the users, they have proven to be great interest to be analysed [14, 38, 48].

This huge amount of content circulating in OSN has attracted the attention of marketing agencies that seek capitalize on the behaviour of clients (current or future) to adjust their online campaigns and even use the social content (number of likes, retweets, etc.) to predict consumer behaviour in the real world [45, 53]. Recently this type of analysis

has jumped into the field of politics in an attempt to predict the results of election campaigns by monitoring the interaction between candidates and voters [9, 27, 47]. The fact that more and more people are posting on the OSN had led to researchers and journalists to believe that a collective feeling is present in the OSN, which can be listened to, captured and analysed [15, 19].

Advances in big data coupled with OSN have also regarded by the scientific communities as an opportunity to mine online social interactions in a more sophisticated way than ever before. Bello-Orgaz et al. [11] coined the term of Social Big Data for the first time as follows: 'Those processes and methods that are designed to provide sensitive and relevant knowledge from social media data sources to any user or company from social media data sources when data source can be characterised by their different formats and contents, their very large size, and the online or streamed generation of information'.

In particular in the case of Twitter, few scholars have identified possible correlations between its activity (such as the frequency, volume of terms…) with general election outcomes. Others are still sceptical about the rigor of these studies and reproducibility of the results [21]. The debate is still incipient and moreover, there is no consensus about the method to apply. Consequently, different approaches have been developed to predict voting outcomes. The first method is directly linked to the volumes and nature of the tweets e.g. retweets, supporters, likes, etc. [24]. The second concerns the techniques of sentiment analysis which compute positive or negative sentiment in the online posts towards candidates or political parties [30]. Finally, more innovative methods analyse the networks of social media users who actively support political parties or candidates [46].

OSN are new means of constructing the reputation of candidates and the strategies of social media campaigns. Nevertheless, in the use context of this information, the question of trustworthiness is paramount [1, 3]. Data credibility may vary depending on the reputation of the data producer or his/her intention to create noise, fake or bad news. The domain of the users also needs to be verified to ensure the tweets can be properly counted for the specific issue studied [56]. This manuscript presents an approach to capture, analyse and predict voting shares of the 2019 President Spanish elections taking into account an analysis of the sentiments expressed by the users of Twitter.

The novelty of our paper is to develop a specific Spanish and political lexicon to classify online messages. To the best of our knowledge, no numerical extended comparison of different inferring methods has been conducted to predict votes using a trained political vocabulary [22]. The experiments carried out use five machine learning techniques for the election candidates. Our best model predicts the winner of the election and the ranking of the remaining candidates.

Considering how the electoral cyberspace is articulated in Spain, several possible spaces could have been considered for the analysis of a political campaign: Facebook, Twitter, Instagram or Pinterest. In this manuscript, Twitter is chosen for two reasons. Firstly, it is an abundant medium for researchers to take the pulse of the public opinion, leveraging the vast volume of content published by its users compared to other social media (200 billion tweets a year).[1] Second, it facilitates easy data collection by providing

---

[1] http://www.dsayce.com/social-media/10-billions-tweets, Last visit 17/4/2020.

API access to the Twitter world unlike other leading OSN like Facebook or Linkedin which do not.

In the race for the presidency, the political context in 2019 was historic. Indeed, five candidates were competing for one of the most important Spanish political elections, including the socialist President Sanchez candidate who was seeking re-election. The other candidates were Rivera for the central-right party, Iglesias for the far-left party, Casado for the right party and Abascal for the far-right party (called "*Vox*"). The outstanding level of participation (75.7%) shows the importance of this scrutiny due to two principal factors.

On one hand, the context of high political divide due to the question of the Catalan referendum for independence and on the other the emergence of the far-right party whose manifesto included controversial issues such as the end of the authorized abortion or a recentralisation of the national governance around Madrid, the capital of Spain. The fact that "Vox" party is competing in this type of election for the first time challenges the standard system of polls forecasting, which are usually based on the results obtained in the previous elections [26]. Voting participation was about 9 points higher than in the previous election and the highest turnout of the twenty-first century. The Presidential election was held under a one-round voting system which took place on April 28th, 2019.

The key contributions of this paper are three-fold:

- To develop a specific framework to classify the sentiment of online users in a context such as a political Presidential election,
- To implement and benchmark five machine learning methods to determine the optimal technique to be used to infer the voting share,
- To predict the election results and the ranking of the candidates.

This remainder of this paper is organized as follows. Next section reviews the literature on the use of social media to infer elections. Then, we explain the methodology used to analyse sentiment from tweets. Consequently, our results are presented and analysed. Finally, we conclude proposing further lines of research to extend and improve this paper.

Related work

OSN are channels for politicians to construct their digital image and reputation [23]. They provide important platforms for the interaction and mobilization of voters before an election [14]. Amongst other online networks, Twitter microblogging allows its users to express their thoughts, feelings and opinions in the form of text. Users can post on a wide variety of issues and it is a space from which enables data extraction on public opinion using linguistic programming models. Then, Machine learning (ML) techniques are commonly used to determine if a message posted by an user expresses a positive or negative sentiment [38, 43, 51]. These methods of analysis are known as sentiment or sentimental analysis (SA).

Le et al. [30] claim that not all the potential voters express their opinion on Twitter and consequently the sample of data collected is not representative of the population.

Grimaldi *et al. J Big Data*   (2020) 7:58

Page 4 of 19

A real challenge resides in the capacity to manage and extract useful knowledge from online media data sources. They add that these networks easily allow the manipulation from propagandists who only add noise to the real message. Indeed, Auletta et al. [7] show how an heuristic is able to add links inside a social network with the objective to promoting a specific candidature. Their results reveal the extraordinary power of online media and the potential risks which emerge from the ability of political or marketing digital experts to control them. This risk represents a great threat as their power can be used to influence undecided voters and consequently affect results.

To improve accuracy in the SA results, Bansal and Srivastava [10] suggest widening the analytical to include emoji sentiments as part of the study. Other studies propose a verification of the consistency of the messages conveyed crossing information at post, user and domain level [1]. The trustworthiness of the social media data is indeed paramount [56]. In this aim, Abu-Salih et al. [1, 3] developed a credibility framework incorporating semantic analysis and temporal factors to measure the domain-based credibility of the users. They further suggest a method to discover influential domain-based users, rank them and identify "super-users" i.e. influencers of a specific online campaign.

Further works have examined if it is possible to infer election results using tweets and ML prediction models [9, 24, 27]. In a seminal work, Tumasjan et al. [49] looked for the margins of error between online and research institutes (traditional channels) taking as example the latest general elections in Germany. More recently, Morris [41] analyses the 2016 U.S. Presidential campaign and compares the results. His conclusions are that the analysis of tweets mentioning a political party could be considered as a plausible reflection of the vote share. Moreover, its predictive power falls close to traditional election polls.

Although a number of techniques and ML algorithms can be used to predict outcomes, this research domain is still incipient. They are based either on counting tweets [17, 27, 35] or on a lexicon-based SA, which is similar to that used in the scope of this paper [10, 26, 29, 39, 42].

The inferring performance depends largely on the capacity to process and analyse users' information. It is also directly linked to the capability to discard user's messages that are classified as spam i.e. unsolicited and repeated junk messages [2–4, 56]. These tweets come usually from bots and have a malicious intention to create rumours and chaos [46]. Shmargad and Sanchez [47] compare the number of followers/enthusiasts of each candidate on Twitter for the 2014 U.S. Senate and the 2016 Congressional elections. They conclude that despite spending less money overall on their campaigns, candidates with greater indirect influence on Twitter (i.e. more likes or retweets), obtain better results than their rivals and have smaller vote gaps than their respective national candidate (i.e. Hillary Clinton or Donald Trump). OSN become a much cheaper alternative than traditional channels (TV, radio, etc.) for politicians with little resources to start a campaign and build a name for themselves.

Prior to the 2016 presidential election, several studies [8, 18] stated that the number of followers (they coined them as indegree) was a sign that revealed the popularity of a user but not necessary related to his/her ability to influence voters. Their study was summarised by the slogan of the 'fallacy of one million followers' which means that

Grimaldi *et al. J Big Data*   (2020) 7:58

Page 5 of 19

users follow others out of politeness or simply become followers of those who also follow them but have hardly read the tweets posted of those they follow.

But, the 2016 race for the White House shows that traditional media and Twitter platforms are critical mechanisms that influence voters [41]. Twitter by conveying candidates' messages about their ideas, policies, and future actions establishes fundamentals that resonate for the public as much as those sent by the traditional media. Consequently, there is growing interest in the creation of textual or linguistic classifiers to determine whether opinions expressed by individuals can be considered as favourable or unfavourable indicators of voting behaviour [9, 14, 48]. The methodological approaches that have been used so far, have been mainly based on a lexicon of specific terms used or pre-tagged texts. Nevertheless, they have not achieved the level of accuracy that those obtained in the press articles written by journalists on topics such as business or consumer goods [45]. In this aim, a specific lexicon for Spanish elections is developed in this paper.

Grimaldi [23] demonstrates that the number of information/conversations/messages circulating on Twitter in the 2019 Spanish Presidential elections grows as the election campaign period progresses, showing peaks in line with relevant events that occur or affect it. The volume of these posts falls after the election voting day. The SA technique needs to adapt to this evolution incorporating for instance new hashtags which were not present at the beginning of the campaign [23]. The fact that campaigns are primarily manifested through Twitter media before standard general channels like TV or radio raises the need to continue investigating the best approach to analyse the data available in these networks and reveal the messages conveyed. This article deals with this controversial line of research. It aims at answering first if it is possible to infer the voting share for each candidate in an election by using Twitter, and secondly, if this media is more reliable than standard public opinion polls in a new political landscape [47].

Methodology

This section presents the different steps of the SA methodology. In particular, in the last section, the methodology used to create a new political lexicon is detailed. The description of the six ML techniques used for SA and to infer the voting intention for each candidate are discussed in the next section.

The SA method is related to Natural Language Processing (NLP), computational linguistics, and text mining [32]. In sentiment classification, which is also presented in this section, the Logistic Regression (LR) is used to determine whether a tweet expresses a positive or negative sentiment. The process of data classification consists of two phases; the training phase along with the prediction phase as shown in Fig. 1.

Data collection

With the Twitter Streaming API,[2] all the posts containing one or more hashtags related to the presidential election are collected every day as follows:
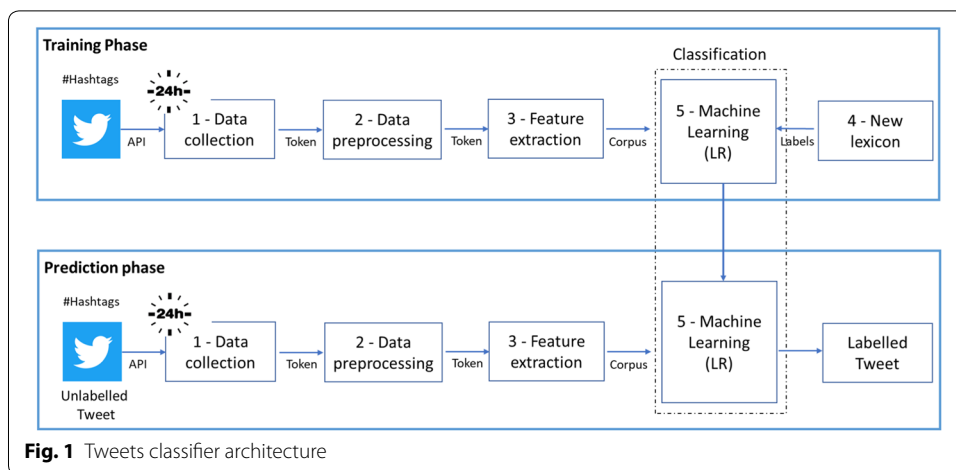
---

[2] https://dev.twitter.com/streaming/overview. Last access on 10-11-2019.

Fig. 1 Tweets classifier architecture

Table 1 Classification of the keywords by political faction

Candidate	Nº Tweets
Pedro Sanchez	320.760
Pablo Iglesias	226.472
Albert Rivera	217.423
Pablo Casado	217.380
Santiago Abascal	188.828

- Election general information=#28A; #28Abril; #Vota; #Vota28A; #Elecciones Generales; #28 Abril; #EspanaVaciada.
- Political party=#ValorSeguro; #Casado; #laEspaña que quieres; #PedroSanchez; #LaHistoriaLaEscribesTu; #PabloIglesias; #VOX; #SantiagoAbascal; #VamosCiudadanos; #AlbertRivera.

The tweets are collected in real time, 24 h after their publication. The aim is to get statistics related to the popularity of the tweet such as (a) the number of shares or retweets, which shows how many users shared the tweet with their followers; (b) the number of likes, which shows how many users linked the tweet (c) the number of retweets which shows how many tweets were re-broadcasted on the media. Between April 12th and April 28th 1.170.000 tweets were posted and Table 1 shows the amount of data collected in terms of users and tweets per candidate.

When working with the Twitter API it was necessary to take into account certain limitations of it. The frequency limit of the interface only allows 450 requests every quarter of hour.

Data preprocessing

The dataset resulting from data collection is a collection of tweets. In order to apply SA techniques, data must go through a four-step process, known as pipeline, in which the input of each step is the output of the previous one.

Grimaldi *et al. J Big Data*   (2020) 7:58

Page 7 of 19

The first step is called the text cleaning. A tweet is a complex object with many properties. But what it is important is the message that the user writes in the form of a text. So, the URLs, emails are deleted and necessary functional words of the Twitter lexicon such as \RT, \HO, \HT, unknown characters are replaced to their closest ASCII variant, using the R Unicode Library. Each document is then transformed into a list of words between blanks or punctuation which is called tokens. During this step, the information about the author of the tweet and the date of publication are lost. But the reference to the object to be able to link the result with the original tweet is kept.

The second step is called the lexical normalization. It removes the functional words which do not have a clear referential semantic, such as articles, pronouns, preposition such as ("the", "an", "and", "of", etc.), numbers or dates. They are usually called as stop words. Uppercase letters are transformed to lowercase. This allows the algorithm to consider identical words such as "Big" and "big" and consequently as two counts of the same word. It deletes also any repeated sequences of letters in the words (such as "greaaaat", "mooorning", etc.) reverting them back to their original English word.

Feature extraction

The next step is the feature extraction applying the term frequency (TF) and Inverse Document Frequency (IDF) R package. The Vector Space Model (VSM) is one of the most commonly used methods for text retrieval and to represent a document as a vector of terms [31, 36]. It consists of transforming each token (i.e. tweet text) into a feature vector in an Euclidean space.: $t = \{t1, t2,... tn\} \in R^s$ where s is the size of the vocabulary which is built by considering unigrams in the collection (i.e. the training test) [33]. The sum of tokens builds a sparse matrix or also called Corpus (C).

New polarity political lexicon

A SA system is highly sensitive to the domain in which the data used to train are extracted. It can obtain poor results if the training dataset is not political [4, 56]. Due to the lack of a sentiment lexicon for non-English languages, the creation of a new polarity lexicon is decided for the Spanish political event issuing from two different sources. The starting point is the dataset created by [40] for Spanish tweets. Then, a random sample of 1.000 tweets is chosen as the training set. In order to label a tweet as either positive, negative or neutral, eight volunteers in group of 2 (i.e. 4 groups) are asked to assign a label for each tweet according to the sentiment they feel it conveys. The evaluators have access to the full tweet, which is tagged according to its global polarity, indicating whether the text expresses a positive, negative or neutral sentiment. To prevent ambiguity, testers have to assess the tone of some tweets after screening the other tweets of the account. Moreover, if there is no agreement inside the group, a third independent volunteer from another group is asked to help to decide. Volunteers agreed on 80% of tweets which support the statement [52] and determined which additional word had to be included in the lexicon. In total 187 words related to this political campaign are added based on this manual testing and inserted in the final vocabulary. These additional words are for instance: "movilización, impulso, progreso, traición, sinverguenza, matando, etc."—English translation: "Mobilization, impulse, progress, treason, shamelessness, killing, etc.).

Machine learning techniques

This section presents a detailed description of the ML techniques used to analyse the collected data, search for patterns and to create estimated voting predictions. In particular, the settings used to train the Machine are explained in detail in the hyper-parameter and evaluation criteria subsections.

The choice of these algorithms is motivated by a review of the existing literature that compares their features in terms of regression method (linear versus non-linear), complexity, performance, accuracy, memory and processing requirement along with time required for training classifier [16, 38].

Logistic regression

The method of estimation of a linear regression consists of determining the coefficients which minimize the sum-of-squared deviations of the observed values of Y from the predicted values based on the model. Two principles distinguish a Logistic Regression (LR) from a linear regression. The first is that the outcome variable in LR is binary or dichotomous. In any regression problem, the key quantity is called the conditional mean and is expressed as 'E (Y |x)' where Y denotes the outcome variable and x denotes a specific value of the independent variable. In the linear regression model, an observation of the outcome variable may be expressed as $y = E(Y|x) + \varepsilon$. The quantity ε is called the error and expresses a deviation from the conditional mean. The most common assumption for a linear regression is that ε follows a normal distribution with mean zero and some variance is constant across levels of the independent variable. In turn, in a logistic regression, the binomial distribution is the statistical distribution of the errors on which the analysis is based [25]. This is the second difference.

The SA system classifies a given vector as either positive, negative or neutral by adding all the number of positive words and subtracting the number of negative words. It assigns only one label class to the tweet. This type of analysis makes sense if each tweet expresses a single opinion. It may seem like a limitation, but in practice it works well since users usually focus on a single topic in each tweet. Surely in other contexts, or if this limitation of message length were not available, it would be necessary to consider more complex systems that allow more granular analysis [12].

Linear discriminant analysis

Linear discriminant analysis (LDA) also known as Fisher Discriminant Analysis is very similar to Logistic regression [55]. It is a method used to find a linear combination of features that separates different classes of objects. LDA searches for those vectors in the underlying space that best discriminate among classes (rather than those that best describe the data). The LDA is optimized when the independent variables are normally distributed. It is indeed an important constraint of the model. Mathematically wise, for all the samples of all classes, two measures are defined (1) and (2):

1. One is called within-class scatter matrix, as given by:

$$S_\omega = \sum_{j=1}^{c} \sum_{i=1}^{N_j} \left(x_i^j - \mu_j \right) \left(x_i^j - \mu_j \right)^T \tag{1}$$

where x_i^j is the ith sample of class j; μ_j is the mean of class j; c is the number of classes and N_j the number of samples in class j.

2. The other is called between-class scatter matrix, as given by:

$$S_b = \sum_{j=1}^{c} \left(\mu_j - \mu \right) \left(\mu_j - \mu \right)^T \tag{2}$$

where μ represents the mean of all classes.

Then, LDA can be formulated as an optimization problem to find a set of linear combinations (with coefficients *w*) that maximizes the ratio of the between-class scattering to the within-class scattering. One way to do this is given by the following generalized eigenvalue problem (3)

$$S_b w = \lambda S_w w. \tag{3}$$

Generally, at most c − 1 generalized eigenvectors are useful to discriminate between c classes.

Classification and Regression Trees

Classification and Regression Trees (CART) is an umbrella term used to refer to decision or regression trees [13]. While the target variable can take a discrete set of values, decision trees are called classification trees. It is a flow-chart structure where each fork represents conjunctions of features that lead to class labels (also called the leaves). Since the branch can be seen as a split in a predictor variable, each end node contains a prediction for the outcome variable. When the target variables can take continuous values (i.e. real numbers such as voting prediction), decision trees are named regression trees. The decision tree continues to expand with additional nodes being repeatedly inserted until the stopping criteria is met. In other words, it stops when the predicted number of iterations is reached, or a reasonable prediction is achieved.

k-Nearest Neighbourhoods

k-Nearest Neighbour (kNN) classification determines the local decision boundaries. The technique assigns any object to the majority class of its k closest neighbours. The main constraint of this method is that the results may change if we change k. KNN is also a 'lazy learner' since all computations are deferred until function evaluation. In this study, the kNN regression is used to estimate continuous variables. This algorithm uses an average of the k nearest neighbours, weighted by the inverse of their distance.

Support Vector Machines

Vapnik [50] firstly introduced Support Vector Machines (SVM) to solve two-class recognition problems. The general idea is to find the decision surface that maximizes the margin between data points, i.e. between classes. A good separation is achieved by the hyperplane that is furthest from the nearest training-data point of any class (so-called functional

margin). In general, the larger the margin is, the lower the generalization error of the classifier is also. In the case of originally linearly non-separable data points, the original data vectors are mapped to higher dimensional space to achieve linear separability again.

Random Forest

The Random Forest (RF) is an algorithm that randomly creates and merges multiple decision trees into one "forest." The goal is not to rely on a single learning model, but rather a collection of decision models to improve accuracy and reduce the variance. The primary difference between this approach and the standard decision tree algorithms is that the root nodes splitting is generated randomly. RF corrects the decision trees recurrent error of overfitting of the training set i.e. it has low bias, but a very high variance. RF generates many predictors, each using regression or classification trees. It reduces instability by averaging multiple decision trees, trained on different parts of the same set of data. Then, it develops a final prediction which usually improves the prediction performance of a standard decision tree.

Hyper-parameter settings

According to Jungherr [28], model performance tends to fluctuate depending on the methods employed even if the impact of the parameters is hard to know. Heredia et al. [24] add that models based on specific counts of tweets that only mention a single party outperform others that combine SAs with volumetric information. Therefore, a correct setting of the hyper-parameters helps to limit negative effects like over- or under-fitting. Respectively in an SVM, kNN or RF method, a higher C, k, or max-depth may cause the model to misclassify less but is much more likely to generate overfit. The Appendix 1 presents the selected settings of hyperparameters. The experiments for this study were carried out using the *fit.models* library in R software. The evaluation criteria for the ML techniques are presented in the next sub-section.

Evaluation criteria

This paper incorporates certain evaluation metrics to validate the efficiency of the proposed model. The metrics MAE and RMSE are used to analyse and compare the performance of ML methods. The Mean Absolute Error (MAE) measures the average magnitude of the errors in a sampling of predictions with absolute differences between prediction and actual observation where all individual differences have equal weight. It is formulated in (4). In turn, the Root Mean Squared Error (RMSE) measures the average magnitude of the error, which is the square root of the average of squared differences between prediction and actual observation. The equation is (5).

$$MAE = \frac{1}{n} \sum_{j=1}^{n} \left| y_j - \widehat{y}_j \right| \tag{4}$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{\left(y_i - \widehat{y}_i \right)^2}{n}}. \tag{5}$$

Since the errors are squared before they are averaged, the main differences between RMSE and MAE is that RMSE gives a relatively high weight to large errors. This means it would be more useful when large errors are particularly undesirable [54].

Jaidka et al. [27] state such consideration is important when minor political parties end up becoming more active on social media platforms than leading parties. In this study, large and small parties are equally active (Table 1), so the five models are benchmarked using MAE results.

Precision, Recall, and F-score are used to measure classification performance. Formulas applied are, respectively (6) and (7) [32]:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

where TP = True positive; TN = True negative; FP = False positive; FN = False negative.

Precision is the division of retrieved instances that are significant. Recall is the fraction of applicable instances that are recovered. In dual classification, recall is known also as sensitivity i.e. the possibility that a relevant document is recovered by the query.

$$F{-}score = 2 * \frac{Precision * Recall}{Precision + Recall}. \tag{8}$$

The F-score balances the use of precision and recall measuring the accuracy of experiments. It is the weighted harmonic mean of the precision and recall (8).

Results

This section presents the main results of the study. The performance of the sentiment classifier and the five inferring models are shown followed by a comparison of them. Finally, a prediction of the votes is presented in the last sub-section using the best ML technique.

### Linear discriminant analysis classifier

For the lexicon-based sentiment classifier, the data are split and the process of cross validation is repeated during k iterations. K is optimized when increased slowly and classifier accuracy is sought. The optimization process (k = 3) is presented in Fig. 2.

The system achieves a macro-averaged F1-score of 92.65% and an accuracy of 92.70% on the test set. These similar scores also confirm there is neither over-, nor under fitting of the model.

The five Machine learning inferring model

The dataset used to predict the voting share of each candidate is composed of the following independent variables:

1. Day tweet volume: the sum of tweets on that day mentioning candidates.
2. Day unique tweet volume: the sum of tweets on that day that only mentions one candidate.

Fig. 2 Optimizing k for the sentiment analysis classifier
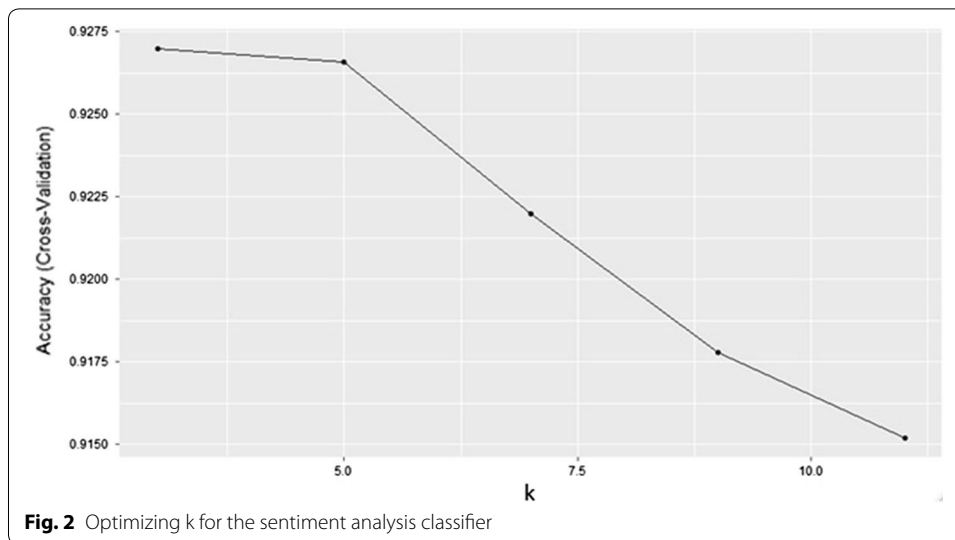
3. Day Twitter user number: the sum of different Twitter users with at least one tweet mentioning a candidate.
4. Day unique Twitter user number: the sum of different Twitter accounts whose posts only mention a candidate.
5. Positive or negative tweet volume: the sum of positive or negative posts that mention a candidate.
6. Positive or negative-based Twitter user number: the sum of different Twitter users with at least one positive or negative post mentioning a candidate.
7. Sentiment score per tweet.

The features are computed in a daily basis. Then they are normalized by applying the moving average smoothing technique over a window of the past 7 days. The polling data is considered as the dependent variable of our model.

Comparison of the results

The voting intention inference is approached as a multiple regression analysis. In this way, several regression models are built to infer the vote of each candidate in the electoral campaign, using the aggregated polling as the output variable of the models. In total, five models are built i.e. one model per candidate. The five different algorithms: LDA, CART, kNN, SVM and RF are then evaluated.

The performance results of the five algorithms are analysed for one candidate: Pedro Sanchez (Fig. 3). The complete results by model and candidate are shown in Table 2. The empirical results demonstrate that kNN with the lowest MAE are the best predictions and outperforms the rest of algorithms on the five models. LDA with the highest MAE has the worst performance.

Inference results from the kNN best model

Using the function predict in the R package, the 2019 Presidential voting shares are shown in Table 3 using the best model (kNN).
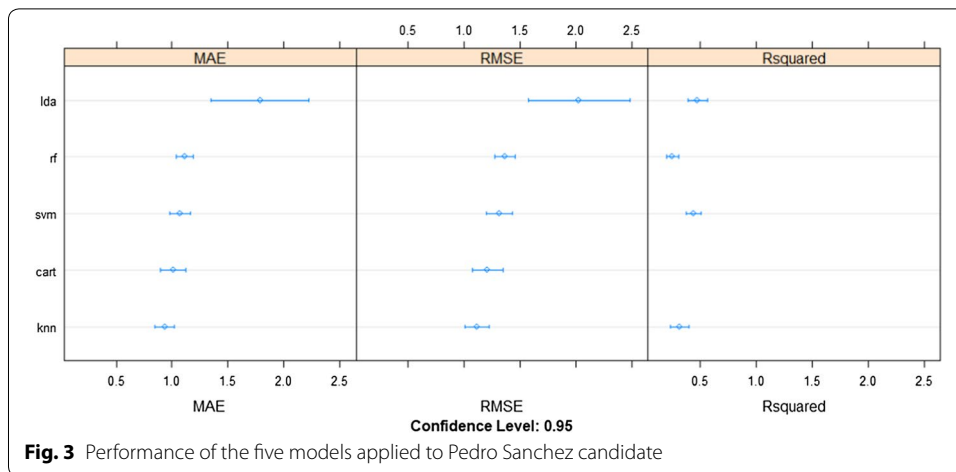
Fig. 3 Performance of the five models applied to Pedro Sanchez candidate

Table 2 Performance of the five models by candidate

MAE	LDA	CART	kNN	SVM	RF
Pedro Sanchez	1.79	1.04	0.95	1.09	1.13
Pablo Iglesias	2.2	1.41	1.2	1.40	1.59
Albert Rivera	3.1	2.55	2.16	2.5	2.64
Pablo Casado	2.8	2.01	1.96	2.1	2.14
Santiago Abascal	1.32	0.67	0.58	0.62	0.76

Table 3 Results and voting inference using kNN method

Candidate	Official election results (in % and in decrescent order) (%)	Polls (difference) (%)		Inference with kNN method (%)	
		Results	Difference with elections	Results	Difference with elections
Pedro Sanchez	28.7	27.3	− 1.4	29.2	+ 0.5
Pablo Casado	16.7	19.4	+ 2.7	19.9	+ 3.2
Albert Rivera	15.9	15.8	− 0.1	14.9	− 1.0
Pablo Iglesias	14.3	14.4	+ 0.1	12.8	− 1.5
Santiago Abascal	10.2	10.2	+ 0.0	10.2	+ 0.0

The first column of Table 3 shows the 2019 Spanish Presidential official results and in the third one the voting intention inference based on kNN best method. As far as the second column is concerned, it presents the last polling results taken 5 days before the election in compliance with Spanish law regulating the presidential election campaigns. The results are discussed in the next section.

Discussion

The significant number of undecided voters at the beginning of the electoral campaign and the political conflict between regions (especially between Catalonia and Madrid) presented us with an interesting challenge in predicting the outcome of the Spanish

Presidential election. The results of this study show the method (a) correctly ranks the candidates and determines the winner (b) gives for the winner of the election (*Sanchez*) a better prediction of voting share than the previous and definitive polls, (c) provides a prediction equivalent as the last poll for *Abascal* candidate, the "surprise" of the scrutiny since his party "*Vox*" was competing for the first time in these elections.

Algorithms comparison with State of Art

Presidential campaign prediction using Twitter data is a much-discussed research topic and the number of studies are still growing [9, 10, 24, 27, 37, 51]. There is, however, no real benchmarking on the performance [22] to identify the best method to determine the winner of the elections or the voting share with the minimum error. Nevertheless, works from other research domains such as Medicine, Engineering or Finance, highlight calibration as a remarkably effective method to obtain better performance [16]. They conclude also LR performs better for smaller data sets and has worst results for larger ones [44]. The inferring results of this study corroborate them. Indeed, LDA with the poorest MAE has the lowest training time required compared to the other algorithms and the data set comprised of 17 days of campaign and 1.170.000 tweets can be considered as large.

Inferring results comparison with State of Art

When predicting the vote rates, Gayo-Avello [22] confirm MAEs are normally used to measuring accuracy. However, they inform that it is difficult to compare MAEs between campaigns. Having said that, a MAE baseline between 1 and 2% can be considered as a good prediction model. Consequently, the MAE of the kNN algorithm positions this paper in the upper quadrant of Twitter prediction performances [22].

Undecided voters

The results of this paper are less accurate than the research institute to estimate the voting share at candidate level: *Casado*, *Rivera* and *Iglesias.* These latter positioned in the centre of the Political Exchequer (Table 3) make up the zone with a large number of undecided voters. Nevertheless, the sum of their votes is 49.6% according to the polls and 47.6% according to the proposed model. The official election results are 46.9%. So, the difference (as a whole) is better for the model compared to the research institute. It demonstrates that our model has been able to better estimate the high volatility during the campaign and the transition of votes between the three blocks: *Sanchez*, the three candidates (*Casado*, *Rivera* and *Iglesias)* and *Abascal.*

'Social Big Data' is an extremely important asset in the predictions of undecided voters. Silva et al. [48] observe a similar migration of votes among blocks of candidates during the Brazilian electoral campaign. To deal with this issue, they introduce a variable called 'voting decision' in their predicting model which indicates the degree of certainty of the voter in relation to his/her declared vote for president. Further lines of investigation can test their method to the present dataset to check if it improves the overall results.

In the following lines, we suggest additional ways to improve this present model.

Tweeter's credibility

The results reveal a gap between what the user writes in a tweet and what he or she really thinks and finally votes, in other words, the online user credibility. Abu-Salih et al. [4] define the construct "Trust" for social media such as the credibility of the content posted by users in a particular domain. Abu-Salih et al. [1] propose an ontology-based domain framework to analyse the semantic of the tweets considering not only the users' messages but also their metadata e.g. the nature and the volume of Retweets and Likes. Their method aims at better understanding users' interests and at highlighting inconsistent behaviour. In other words, it evaluates tweets credibility in a specific domain. Their approach is unique since most of the studies in the past had neglected the domain level of the message along with the factor of time [2–4]. Further study should analyse if including this framework could improve the inferring results by discarding inconsistent tweets injected in the prediction algorithms. Nevertheless, the outcome could be limited since the tweets contain hashtags which were carefully selected to ensure they were strictly related to the Spanish Presidential election domain analysed (e.g. #28A, #28Abril, #Elecciones Generales).

Spam detection

Many presidential candidates hire companies to create robot accounts to follow them. These anomalous users or spammers post tweets supporting them. Those messages and Twitter accounts don't represent the general public opinions and have to be discarded. Their objective is to create positive or negative rumours. In this manuscript, detection of spams or irrelevant posts is realized by implementing a ML supervised method, a RF algorithm using the following features: spammer users previously identified, age of the user account, retweet rate and number of favourites. But, more recently, Abu-Salih et al. [3] include a time-based semantic analysis at domain and user levels. Their technique provides a solution to rank tweeters' credibility. Consequently, further work could analyse if this novel approach improves the results of the cleansing phase to detect untrustworthy users i.e. spammers.

Computational time

In ML techniques, the splitting of the data is performed in a stratified way and the process of cross validation is repeated during k iterations. This method is more precise but could have the disadvantage of slowing down the ML computational times. Ahmadvand et al. [6] suggest the sampling of the computing techniques as an alternative. They add that this usually generates the desired quality of result when resources such as time, cost or energy are limited. Another option resides on progressive computation i.e. when large data space is pruned progressively to look for the results [5]. The processing times were affordable, but this method could be used if scholars decide to extend the time period of data collection or the number of hashtags and get a larger dataset.

Conclusion and future works

This study has achieved the objectives to develop a fast, cheap and reliable tool for understanding online public sentiment and predicting Spanish Presidential election results. First, it develops a method which correctly ranks the candidates and determines the winner (*Sanchez*). Second, it benchmarks five different ML inferring algorithms and the best model assigns a better prediction of votes to the elected than the official research institutes. Finally, it builds a political lexicon-based framework to measure the sentiments of online users.

For digital marketers and political campaigners, it represents an important tool along with the traditional and standard surveys to gauge the strategy of their candidate and correlate it with potential voting shares. In this line, the dataset was created by mining Twitter for the last 17 days of the campaign; however, future works could extend it and create an automated framework which collects data over a longer period. Indeed, election results predicting is a continuous process and requires analysis over months or years. So, while developing predictive models from social media, time could be an important factor in assessing the sustaining preference for a political party. It could also offset possible campaign incidents such as when a candidate's popularity suddenly skyrockets.

Scientifically speaking, this study has implemented and compared five machine learning methods to determine the optimal technique which was used then to infer the voting share. It obtains an error of estimation situated in the upper quadrant of Twitter prediction performances. It has also leveraged a specific lexicon for Spanish political context. However, we believe two additional lines of research could improve this work. On one hand, future studies should analyse and propose weighting factors to include the 'inaudible' voice, i.e. the part of the population who does not express their opinion using social media tools, in the statistical balance. On the other hand, they should find a way to process demographic data (e.g. sex, age, and geographic location) in the method so as to be at par with the statistical sampling methods. The final objective is to reduce the response bias and obtain a more representative vision of the electorate.

Author details
[1] Department of Strategy, Entrepreneurship and Innovation, La Salle Campus, Ramon Llull University, Sant Joan de La Salle 42, 08022 Barcelona, Spain. [2] Universidad Icesi, Calle 18 No. 122-135 Pance, Cali, Colombia.

Appendix 1: Hyper-parameter settings for ML methods

Linear discriminant analysis (LDA)		
Penalty	The method of penalization of the coefficients of non-contributing variables	Lasso (L1) performs feature selection as it shrinks the less important feature's coefficient to zero
C	It is the inverse of the regularization term (1/lambda). It tells the model how much large parameters are penalized	0.001
Class_weight	It permits to place greater emphasis on a class	Balanced weights. All weights are 1
Classification and Regression Trees (CART)		
Fit intercept	It specifies whether the intercept of the model should be calculated or set to zero	True. The model calculates the intercept.
Normalize	It specifies whether to normalize the data for the model using the L2 norm	True
k-Nearest Neighbours (kNN)		
k	It determines the number of neighbours used when calculating the nearest-neighbours algorithm	3 for the classifier 10 for the inferring model
P	It is the power metric when calculating the Minkowski metric	2
Support Vector Machines (SVM)		
C	It is the inverse of the regularization term (1/lambda). It tells the model how much large parameters are penalized	0.001
Class_weight	It sets the parameter of class i to be class weight[i] *C allowing to place greater emphasis on a specific class	Balanced weights. All weights are 1
Random Forest (RF)		
N_estimators	It sets the number of decision trees to be used in the forest	100
Max_depth	It sets the max depth of the tree	30
Min_samples_split	It is the minimum number of samples needed before a split (differentiation) is made in an internal node	10
Min_samples_leaf	It is the minimum number of samples needed to create a leaf (decision) node.	1
Max_features	It sets the number of features to consider for the best node split	Auto

Grimaldi *et al. J Big Data*   (2020) 7:58

Page 18 of 19

References

1. Abu-Salih B, Bremie B, Wongthongtham P, Duan K, Issa T, Chan KY. Social credibility incorporating semantic analysis and machine learning: a survey of the state-of-the-art and future research directions. Adv Intell Syst Comput. 2019;927:887–96. https://doi.org/10.1007/978-3-030-15035-8_87.
2. Abu-Salih B, Chan KY, Al-Kadi O, Al-Tawil M, Wongthongtham P, Issa T, et al. Time-aware domain-based social influence prediction. J Big Data. 2020;7(1):10. https://doi.org/10.1186/s40537-020-0283-3.
3. Abu-Salih B, Wongthongtham P, Chan KY, Zhu D. CredSaT: credibility ranking of users in big social data incorporating semantic analysis and temporal factor. J Inf Sci. 2019;45(2):259–80. https://doi.org/10.1177/0165551518790424.
4. Abu-Salih B, Wongthongtham P, Yan Kit C. Twitter mining for ontology-based domain discovery incorporating machine learning. J Knowl Manage. 2018;22(5):949–81. https://doi.org/10.1108/JKM-11-2016-0489.
5. Ahmadvand H, Goudarzi M. Using data variety for efficient progressive big data processing in warehouse-scale computers. IEEE Comput Archit Lett. 2017;16(2):166–9. https://doi.org/10.1109/LCA.2016.2636293.
6. Ahmadvand H, Goudarzi M, Foroutan F. Gapprox: using Gallup approach for approximation in Big Data processing. J Big Data. 2019;6(1):20. https://doi.org/10.1186/s40537-019-0185-4.
7. Auletta V, Ferraioli D, Savarese V. Manipulating an election in social networks through link addition. J Ambient Intell Humaniz Comput. 2020. https://doi.org/10.1007/s12652-019-01669-5.
8. Avnit A. The million followers fallacy. Internet Draft, Pravda Media. 2009. http://tinyurl.com/nshcjg.
9. Awais M, Hassan SU, Ahmed A. Leveraging big data for politics: predicting general election of Pakistan using a novel rigged model. J Ambient Intell Humaniz Comput. 2019. https://doi.org/10.1007/s12652-019-01378-z.
10. Bansal B, Srivastava S. Lexicon-based Twitter sentiment analysis for vote share prediction using emoji and N-gram features. Int J Web Based Commun. 2019;15(1):85–99. https://doi.org/10.1504/IJWBC.2019.098693.
11. Bello-Orgaz G, Jung JJ, Camacho D. Social big data: recent achievements and new challenges. Inf Fusion. 2016;28:45–59. https://doi.org/10.1016/j.inffus.2015.08.005.
12. Bode L, Dalrymple KE. Politics in 140 characters or less: campaign communication, network interaction, and political participation on twitter. J Polit Mark. 2016;15(4):311–32. https://doi.org/10.1080/15377857.2014.959686.
13. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Belmont: Wadsworth and Brooks; 1984.
14. Buccoliero L, Bellio E, Crestini G, Arkoudas A. Twitter and politics: evidence from the US presidential elections 2016. J Mark Commun. 2020;26(1):88–114. https://doi.org/10.1080/13527266.2018.1504228.
15. Budiharto W, Meiliana M. Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis. J Big Data. 2018;5(1):1–10. https://doi.org/10.1186/s40537-018-0164-1.
16. Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. In: ACM International conference proceeding series, vol. 148. 2006. pp. 161–8. https://doi.org/10.1145/1143844.1143865.
17. Cerón-Guzmán JA, León-Guzmán E. A sentiment analysis system of Spanish tweets and its application in Colombia 2014 presidential election. In: Proceedings—2016 IEEE international conferences on Big Data and Cloud Computing, BDCloud 2016, Social Computing and Networking, SocialCom 2016 and Sustainable Computing and Communications, SustainCom. 2016. pp. 250–7. https://doi.org/10.1109/BDCloud-SocialCom-SustainCom.2016.47.
18. Cha M, Gummadi KP. Measuring user influence in Twitter: the million follower fallacy. 2010. http://en.scientific commons.org/58470236.
19. Cury RM. Oscillation of tweet sentiments in the election of João Doria Jr. for Mayor. J Big Data. 2019;6(1):1–15. https://doi.org/10.1186/s40537-019-0208-1.
20. Dietrich BJ, Juelich CL. When presidential candidates voice party issues, does Twitter listen? J Elect Public Opin Parties. 2018;28(2):208–24. https://doi.org/10.1080/17457289.2018.1441847.
21. Gayo-Avello D. No, you cannot predict elections with twitter. IEEE Internet Comput. 2012;16(6):91–4. https://doi.org/10.1109/MIC.2012.137.
22. Gayo-Avello D. A meta-analysis of state-of-the-art electoral prediction from twitter data. Soc Sci Comput Rev. 2013;31:649–79. https://doi.org/10.1177/0894439313493979.
23. Grimaldi D. Can we analyse political discourse using Twitter? Evidence from Spanish 2019 presidential election. Soc Netw Anal Mining. 2019. https://doi.org/10.1007/s13278-019-0594-6.
24. Heredia B, Prusa JD, Khoshgoftaar TM. Social media for polling and predicting United States election outcome. Soc Netw Anal Mining. 2018. https://doi.org/10.1007/s13278-018-0525-y.
25. Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. Hoboken: Wiley; 2013.
26. Huberty M. Can we vote with our tweet? On the perennial difficulty of election forecasting with social media. Int J Forecast. 2015;31(3):992–1007. https://doi.org/10.1016/j.ijforecast.2014.08.005.
27. Jaidka K, Ahmed S, Skoric M, Hilbert M. Predicting elections from social media: a three-country, three-method comparative study. Asian J Commun. 2019;29(3):252–73. https://doi.org/10.1080/01292986.2018.1453849.
28. Jungherr A. Twitter use in election campaigns: a systematic literature review. J Inf Technol Polit. 2016;13(1):72–91. https://doi.org/10.1080/19331681.2015.1132401.
29. Jungherr A, Jürgens P, Schoen H. 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;30(2):229–34. https://doi.org/10.1177/0894439311404119.
30. Le HT, Boynton GR, Mejova Y, Shafiq Z, Srinivasan P. Revisiting The American Voter on Twitter. 2017. pp. 4507–19. https://doi.org/10.1145/3025453.3025543.
31. Letsche TA, Berry MW. Large-scale information retrieval with latent semantic indexing. Inf Sci. 1997;100(1–4):105–37. https://doi.org/10.1016/S0020-0255(97)00044-3.
32. Manning CD, Schütze H, Weikurn G. Foundations of statistical natural language processing. SIGMOD Rec. 2002;31(3):37–8. https://doi.org/10.1145/601858.601867.
33. Manning C, Raghavan P. Introduction to information retrieval. Comput Linguist. 2009. https://doi.org/10.1162/coli.2009.35.2.307.

34. Marozzo F, Bessi A. Analyzing polarization of social media users and news sites during political campaigns. Soc Netw Anal Mining. 2018;8(1):61. https://doi.org/10.1007/s13278-017-0479-5.

35. Metaxas PT, Mustafaraj E, Gayo-Avello D. How (not) to predict elections. In: 2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing, Boston, MA, 2011. pp. 165–171. https://doi.org/10.1109/PASSAT/SocialCom.2011.98.

36. McCarey F, Cinnéide MÓ, Kushmerick N. Recommending library methods: an evaluation of the Vector Space Model (VSM) and Latent Semantic Indexing (LSI). Lecture notes in computer science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 4039 LNCS. 2006. pp. 217–30. https://doi.org/10.1007/11763864_16.

37. McGregor SC, Mourão RR, Molyneux L. Twitter as a tool for and object of political and electoral activity: considering electoral context and variance among actors. J Inf Technol Polit. 2017;14(2):154–67. https://doi.org/10.1080/19331681.2017.1308289.

38. Mehta RP, Sanghvi MA, Shah DK, Singh A. Sentiment analysis of tweets using supervised learning algorithms. Adv Intell Syst Comput. 2020;1045:323–38. https://doi.org/10.1007/978-981-15-0029-9_26.

39. Miura Y, Sakaki S, Hattori K, Ohkuma T. TeamX: a sentiment analyzer with enhanced lexicon mapping and weighting scheme for unbalanced data, (SemEval). 2015. pp. 628–32. https://doi.org/10.3115/v1/s14-2111.

40. Molina-González MD, Martínez-Cámara E, Martín-Valdivia MT, Perea-Ortega JM. Semantic orientation for polarity classification in Spanish reviews. Expert Syst Appl. 2013;40(18):7250–7. https://doi.org/10.1016/j.eswa.2013.06.076.

41. Morris DS. 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. 2018;36(4):456–68. https://doi.org/10.1177/0894439317721441.

42. O'Connor B, Balasubramanyan, Routledge BR, Smith NA. From tweets to polls: linking text sentiment to public opinion time series. In: Proceedings of the fourth international AAAI conference on weblogs and social media. 2010.

43. Patel NV, Chhinkaniwala H. Investigating machine learning techniques for user sentiment analysis. Int J Decis Support Syst Technol. 2019;11(3):1–12. https://doi.org/10.4018/IJDSST.2019070101.

44. Perlich C, Simonoff JS. Perlich_03a.pdf. 2003; 4:211–55. https://doi.org/10.1162/153244304322972694.

45. Rathor AS, Agarwal A, Dimri P. Comparative study of machine learning approaches for amazon reviews. Procedia Comput Sci. 2018;132:1552–61. https://doi.org/10.1016/j.procs.2018.05.119.

46. Shin J, Jian L, Driscoll K, Bar F. Political rumoring on Twitter during the 2012 US presidential election: rumor diffusion and correction. New Media Soc. 2017;19(8):1214–35. https://doi.org/10.1177/1461444816634054.

47. Shmargad Y, Sanchez L. Social media influence and electoral competition. Soc Sci Comput Rev. 2020. https://doi.org/10.1177/0894439320906803.

48. Silva G, Costa M, Drummond A, Weigang L. Predictive model for Brazilian presidential election based on analysis of social media. Adv Intell Syst Comput. 2020;1075:46–53. https://doi.org/10.1007/978-3-030-32591-6_5.

49. Tumasjan A, Sprenger TO, Sandner PG, Welpe IM. Election forecasts with Twitter: how 140 characters reflect the political landscape. Soc Sci Comput Rev. 2011;29(4):402–18. https://doi.org/10.1177/0894439310386557.

50. Vapnik VN. The nature of statistical learning theory. New York: Springer; 1995.

51. Verma P, Khanday AM, Rabani ST, Mir MH, Jamwal S. Twitter sentiment analysis on Indian government project using R. Int J Recent Technol Eng. 2019;8(3):8338–41. https://doi.org/10.35940/ijrte.C6612.098319.

52. Villena J, García J, Martínez E, Jiménez S. TASS 2014—the challenge of aspect-based sentiment analysis. Procesamiento de Lenguaje Nat. 2015;54:61–8.

53. Volkova S, Bachrach Y, Armstrong M, Sharma V. Inferring latent user properties from texts published in social media. In: Proceedings of the twenty-ninth conference on artificial intelligence (AAAI). 2015. pp. 4296–7.

54. Wang W, Lu Y. Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model. In: IOP conference series: materials science and engineering, vol. 324, No. 1. 2018. https://doi.org/10.1088/1757-899X/324/1/012049.

55. Welling M. Fisher linear discriminant analysis max. In: 2007 9th international symposium on signal processing and its applications, ISSPA 2007, Proceedings, (2). 2007. https://www.ics.uci.edu/~welling/teaching/273ASpring09/Fisher-LDA.pdf.

56. Wongthongtham P, Salih BA. Ontology-based approach for identifying the credibility domain in social Big Data. J Organ Comput Electron Commer. 2018;28(4):354–77. https://doi.org/10.1080/10919392.2018.1517481.

Publisher's Note