



# A survey on the use of data and opinion mining in social media to political electoral outcomes prediction

Jéssica S. Santos<sup>1</sup> · Flavia Bernardini<sup>1</sup> · Aline Paes<sup>1</sup>

Received: 11 July 2021 / Revised: 28 September 2021 / Accepted: 1 October 2021 / Published online: 20 October 2021  
© The Author(s), under exclusive licence to Springer-Verlag GmbH Austria, part of Springer Nature 2021

## Abstract

Election polls are the de facto mechanisms to predict political outcomes. Traditionally, they are conducted with personal interviews and questionnaires. This process is costly and time consuming, demanding the development of alternative approaches faster and less expensive. On the other hand, social media emerge as important tools for people to express their opinions about candidates in electoral scenarios. In this context, there is an increasing number of election prediction approaches using social media and opinion mining, modeling this problem in different ways. In this work, we present a survey on approaches to election predictions and discuss many possibilities of decisions in the general process of constructing solutions to this end, including the quantity of collected data, the specific social media used, the collection period, the algorithms and prediction approaches adopted, among others aspects. Our overview allowed us to identify the main factors that should be considered when predicting elections outcomes supported by social media content, as well as the main open issues and limitations of the approaches found in the literature for data science communities. In brief, the main challenges that we have found include but are not limited to: labeling data reliably during the short period of electoral campaigns, absence of a robust methodology to collect and analyze data, non-availability of domain (labeled) datasets, a lack of a pattern to evaluate the obtained results and exploration of new machine learning algorithms and methods for tackling the peculiarities of this scenario.

**Keywords** Election prediction · Election forecast · Election outcomes · Social media · Sentiment analysis · Machine learning for opinion mining

## 1 Introduction

Elections are vital elements for democracy, as they allow citizens to participate in the process of choosing their government representatives, taking into account their will (Katz 1997). Choosing a political candidate to vote for can be a challenging task since it means selecting proposals and policies, advocated by several political parties that may or not share the same ideals of the citizens (Forsythe et al. 1993). On the other hand, the election process is also challenging for the candidates because they need to focus on the

topics that matter to most of the citizens, and, at the same time, they need to clearly and concisely communicate their ideas with the citizens to gain their votes. In this way, in democratic systems, *election polls* play an essential role, as they can measure voting intention (Dwi Prasetyo and Hauff 2015), and their results can affect election outcomes (Forsythe et al. 1993), by influencing people that have not decided yet who they will vote for Rothschild and Malhotra (2014). Additionally, election polls can be used by the candidates and their parties to adjust their campaigns and better communicate proposals (Dwi Prasetyo and Hauff 2015; You et al. 2015).

The traditional way of predicting election outcomes is based on opinion surveys that include face-to-face or phone interviews and questionnaires. These polls involve people from different regions that have different profiles, such as people that live in urban or rural zones, people with different ethnicity, age, gender, among other features. However, interviewing people from different regions is an expensive and time-consuming task (Wang et al. 2015). Additionally,

✉ Jéssica S. Santos  
jessicasoares@id.uff.br

Flavia Bernardini  
fbernardini@ic.uff.br

Aline Paes  
alinepaes@ic.uff.br

<sup>1</sup> Institute of Computing, Universidade Federal Fluminense, Niterói, Rio de Janeiro, Brazil

this task requires time and human efforts to collect data (opinions) across the nation/state (Ibrahim et al. 2015), and in some cases, traditional polls were not able to predict election outcomes correctly, as the famous cases of the 2016 US Presidential Elections (Zeedan 2019; Breur 2016) and the 2015 UK General Elections (Castelvecchi 2017; Sturgis et al. 2018).

Taking into account issues such as time demanded, monetary costs and human efforts required by the traditional election polls (Wang et al. 2015), several approaches have proposed ways to predict voting intention by relying on natural language processing techniques and by applying machine learning in the data collected from social media. As pointed out by Bovet et al. (2018), mining opinions to identify trends based on data collected from the Internet is one of the main goals of the Big Data era (Sagiroglu and Sinanc 2013). Additionally, the growth of the users in social media makes the virtual community look like the real community (Srivastava et al. 2015), what would allow data analysis using social media. For these reasons, social media has been pointed out as a new way of collecting data that can be used to forecast future outcomes in the real world (Burnap et al. 2016; Asur and Huberman 2010). In this context, there is an increasing number of approaches in recent years that use data from social media in order to predict political election results. However, a decade ago, Gayo-Avello (2011, 2012) pointed out that there are many challenging issues in regard to predicting elections outcomes, and it may be even not possible to make these predictions. According to the author, the main problems when trying to predict presidential elections of USA in 2008 were as follows: (i) Big-data fallacy, as having a huge amount of data does not mean the sample is statistically representative of the overall population; (ii) Demographic bias, since users of social media tend to be relatively young; (iii) Naïve sentiment analysis, as some applications might achieve reasonable results by counting topic frequency or using simple approaches to sentiment detection but dealing with political texts may be especially difficult. The work by Santos et al. (2021) reinforces this idea by presenting an exploratory study of the sentiment analysis process of tweets in Portuguese—our mother tongue—about the 2018 Brazilian presidential elections. Such work shows a high level of divergence between labels obtained with automatic labeling strategies and labels obtained through manual labeling using crowdsourcing; (iv) Silence speaks volumes, as nonresponses often play a more important role than collected data; and (v) Past positive results do not guarantee generalization, in the sense that researchers should always be aware of the file-drawer effect and carefully evaluate positive reports before assuming the reported methods are straightforwardly applicable to any similar scenario with identical results.

On the other hand, in the last decade, many papers have presented applications of data and opinion mining technologies, trying to investigate other methods that improve election prediction. Beyond what was pointed out by Gayo-Avello (2011, 2012), many papers also discuss the many peculiarities in electoral scenario has, according to our literature review. An example is the *nature of dispute*, inherent in political elections. While for other domains, such as product and movie reviews, we can make an analysis of opinions based on the sentiment polarity of the documents mentioning a certain product/movie; in the electoral scenario, it is important to find out to whom the user's feeling is directed to. In some cases, the user mentions more than one candidate in the same social media post, supporting one candidate and rejecting another. Additionally, while terms that denote positive/negative sentiment in domains such as movies and products are words that appear in general domains/dictionaries, words that denote sentiment in the election domain many times are *specific terms related to the given election*, as hashtags that combine support messages with candidate names or campaign slogans. Another particular characteristic of the electoral domain is its *dynamic nature*, where the vocabulary changes too fast according to electoral events such as debates, scandals, and public speeches (Mahendiran et al. 2014; Calais Guerra et al. 2011). *Spam and fake news* are factors that should be specially considered in the electoral domain, where a lot of content is posted by bot users (Woolley 2016). Also, high levels of *sarcasm, irony and hate speech*, which are usual elements in social media, are intensified when it comes to political elections (Liu 2020; Gao et al. 2017; Woolley 2016; Okeowo 2016). Those characteristics add noise and can confuse prediction algorithms. According to Liu (2020), the complexity of the topics and sentiment expressions of political texts explain why sentiment analysis commercial systems can achieve good results when applied to analyze opinions about services and products but cannot achieve good results on political opinions. In addition to the aforementioned issues, another critical challenge is the *short time for labeling* electorate opinions (dos Santos et al. 2019). This is due to most of the predictive techniques adopted are based on supervised machine learning algorithms that require labeled data and there is not enough time to manually annotate electorate opinions reliably, during the short period of campaigns. Considering all these peculiarities, it is not easy for software developers and data scientists to make decisions of the process of opinion mining when constructing solutions for elections outcomes predictions based on social media data analysis. Also, from the trustworthy artificial intelligence (AI) point of view, registering all decisions for guaranteeing traceability, accountability and transparency when using AI is very important in decision driven scenarios (Janssen et al. 2020),

specially in very important domains to society such as electoral scenarios.

In this way, we present in this paper a survey on the use of data and opinion mining technologies on social media to political election outcomes prediction, aiming at understanding the existing approaches for these predictions and how data and opinion mining methods and techniques are used. To this end, we conducted a systematic literature review (SLR) to analyze how data and opinion mining on social media data can be efficiently adopted to forecast election outcomes.

## 1.1 SLR methodology

In order to present an overview of the use of technologies of data and opinion mining for election outcomes predictions, we conducted a Systematic Literature Review, focusing on the following research questions:

- **Q1:** What are the main approaches for predicting election outcomes by using social media?
- **Q2:** What are the main data science limitations of the approaches that collect social media data in order to predict elections?
- **Q3:** What are the possible lines for future research on election prediction using social media from the AI point of view?

We conducted our search in the IEEE, ACM, Scopus and Science Direct digital libraries using the following search string:

```
("election prediction" OR "election forecast") AND ("social media" OR "Twitter").
```

This search string was executed on August of 2020, which returned a total of 242 works. We filtered the papers published from 2014, resulting in 207 works to be analyzed. When analyzing the abstracts, we considered the following inclusion criteria: (I1) Papers that propose methods to predict *election outcomes* using *social media* and (I2) Papers that apply existing data and opinion methods for election outcomes prediction based on social media data. Our exclusion criteria were as follows:

- (E1) papers predicting election outcomes not using social media posts. For instance, Li et al. (2017) presented a method to predict elections outcomes considering the results of previous elections and questionnaires, and Garcia et al. (2018) predicted election outcomes of the 2016 Brazilian municipal elections relying on comments extracted from news websites instead of social media posts;

- (E2) works analyzing some aspects of electoral data extracted from social media but do not predict election outcomes. For instance, Sokolova and Perez (2018) analyzed data from Twitter to find out key topics and influencers for the left and right wings for the 2017 French presidential elections; dos Santos et al. (2019) investigated the usage of sentiment analysis for datasets from several domains to predict people sentiment toward the 2018 Brazilian presidential elections; Idan and Feigenbaum (2019) adopted a Bayesian network to predict the voting behavior of a given Facebook user in relation to the US 2016 presidential elections based on his Facebook profile; and Di Giovanni et al. (2018) analyzed tweets posted by Italian deputies to discover the most mentioned topics by political alignment. All of them analyze aspects related to elections but they do not try to predict election outcomes. It is worth mentioning that, although they are out of our scope, they can inspire in the future how to tackle all the issues that raise on data and opinion process for predicting election outcomes using social media;
- (E3) Papers not written in English.

Finally, after the process of filtering, removing the duplicated and unrelated papers and applying the inclusion and exclusion criteria on the 207 papers, we ended up with 53 works. We observed that some of these works use similar strategies for forecasting elections. We categorize the majority of the works into four approaches for elections outcomes predictions, namely *Counting-Based Approach*, *Political Alignment Approach*, *Event Detection Approach* and *Popularity-Based Approach*. The works that could not be categorized in these categories are described in a separate section called *Other Works*. The main differences between this survey and related surveys are based on their scope, focus, and methodological aspects, as illustrated in Table 1.

Skoric et al. (2020), for example, presents an analysis about the influence of several contextual variables such as the democracy score, electoral system type, media freedom and Internet penetration, when predicting election results based on social media. Koli and Ahmed (2019) investigate if factors such as the literacy rate of the country and Internet penetration may contribute to successful sentiment analysis predictions using social media. While Chauhan et al. (2021) and Koli and Ahmed (2019) focus only on the analysis of election prediction approaches that use sentiment analysis, the work by Singh and Sawhney (2018) is even more restricted as it focuses on sentiment analysis approaches and Twitter, specifically. Finally, dos Santos Brito et al. (2021) presents a research closer to ours, but regarding opinion mining tasks, they only covered a few aspects related to data collection. Therefore, differently from the aforementioned surveys, ours focus on the opinion mining process and all its

**Table 1** Methodology, focus, and scope

Paper	Systematic review protocol	Focus	Opinion mining tasks				
			Data collection	Data labeling	Data pre-processing	Demographic info	ML algorithms
Skoric et al. (2020))	✓	Contextual variables					
Chauhan et al. (2021)		Sentiment analysis					✓
dos Santos Brito et al. (2021)	✓	Opinion mining	✓				
Singh and Sawhney (2018)		Sentiment analysis	✓				
Koli and Ahmed (2019)		Sentiment analysis					✓
Our paper	✓	Opinion mining	✓	✓	✓	✓	✓

respective tasks, namely data collection, data labeling, data preprocessing, demographic information aspects, machine learning algorithms, and prediction approach. Since our research is not restricted to sentiment analysis approaches, we provided an approach categorization to group works that use similar strategies to predict elections based on social media. Based on that, we were able to identify the main limitations and challenges for data science in this scenario, and point out open issues from the AI point of view.

## 1.2 Contributions and survey structure

In face of the challenges previously described in regard to political elections outcomes predictions, the main contributions of our work are as follows:

- we summarize and categorize the surveyed works regarding the prediction approaches that they follow (Sect. 3);
- we analyze the characteristics of the selected works according to the process that emerged to us, which extends a more general data and opinion mining process, composed by the tasks: collect electoral opinions from social media (i.e., choose the social media to be used, quantity of data collected, keywords used to collect data and collection period); clean and label data; analyze demographic aspects; select prediction approach; choose machine learning algorithms to be used; and forecast election outcomes. We discuss the very different decisions of the authors, as described in Sect. 4. These specific decisions are important to be identified for each domain for (i) helping data scientists and software engineers in complex scenarios like elections outcome predictions; and (ii) helping the community to enhance trustworthy AI, as there is a need for registering all the decisions to construct solutions for guaranteeing aspects of transparency, auditability and replicability Janssen et al. (2020), specially when considering important domains for society, such as political elections scenarios;

- we present a discussion about the surveyed approaches, identify gaps and limitations in the current literature and point out directions for future research (Sect. 5).

The target audience for this research are mainly people who want to use data and opinion mining technologies to predict elections outcomes using social media. We believe that the content presented in Sects. 3, 4, 5 and 6 can serve as a guide for those who want to explore this research topic. The structure of our work is illustrated in Fig. 1, which is organized as follows: Sect. 2 introduces the basic concepts needed for understanding this paper. Section 3 summarizes the surveyed works, grouping them into different prediction approaches. The ones we could not categorize are described in a subsection named Other Works. Section 4 presents the general data and opinion mining process that we identified for predicting electoral outcomes based on social media data. We took into account the several tasks of this process and pointed out the main decisions made by each work in each one of the tasks. Section 5 presents a discussion about our findings, including the main limitations and gaps, and lines for future directions. Finally, Sect. 6 presents our conclusions.

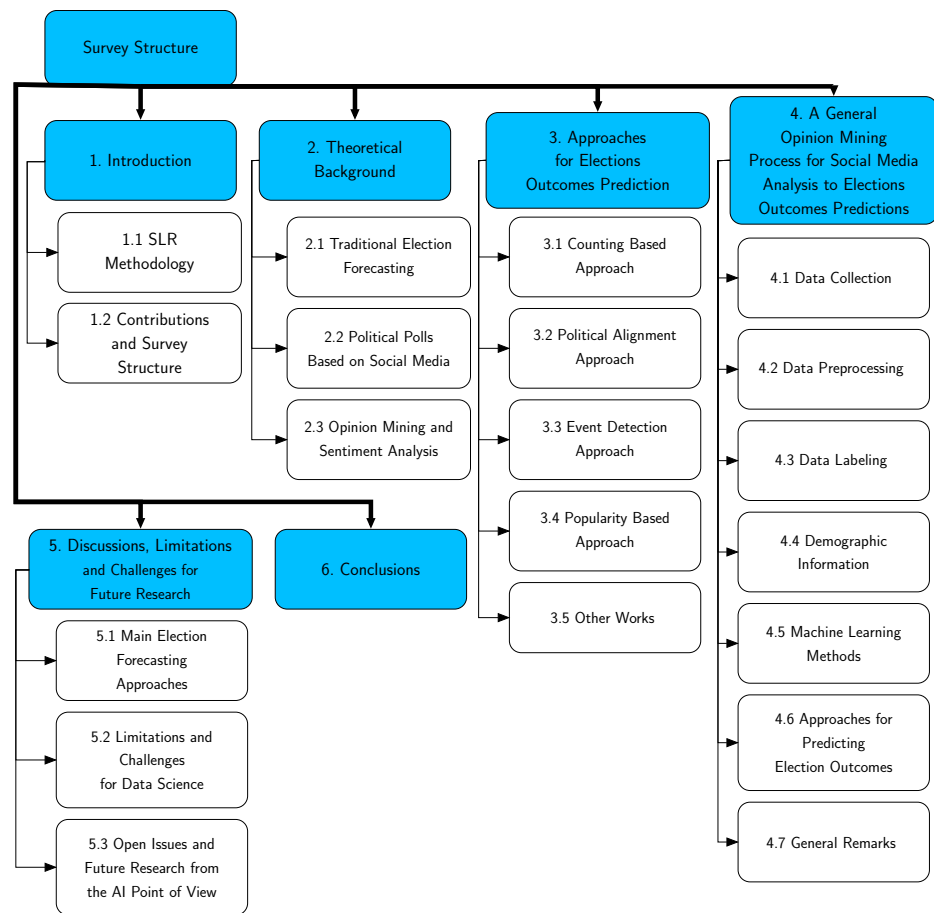
## 2 Theoretical background

This section introduces concepts that are used along of this paper, such as the traditional methods for forecasting election results, how social media can be used to collect opinions and infer voting intention, and the sentiment analysis concept, which is a usual method to automatically extract the polarity of texts, aiding the task of inferring popular opinion.

### 2.1 Traditional election forecasting

The outcomes of political elections are typically attempted to be predicted based on representative opinion polling, which creates a correspondence between the electorate's demographic composition and the population interviewed

Fig. 1 Survey structure



by the survey organizations. Besides aiming at predicting which candidate will be elected, these polls can also be used to infer political trends (Dwi Prasetyo and Hauff 2015). Representative polls are conducted by interviewing a sample of individuals that belong to a particular target population, asking who they are going to vote for Wang et al. (2015).

In order to define the target population, several demographic variables are considered, such as age, sex, state, race, income, education, ideology, and party (Gelman and King 1993). Election polls conducted around the same time by different survey organizations can present high variance in their results due to non-responses, faulty processing, and sample issues. For this reason, their results can be combined in order to reduce measurement error (Graefe 2014). Although most of the time, this poll mechanism can correctly predict the results of elections, there are some cases when the traditional polls were not successful, which is the case for instance, of the 2015 UK general elections, as mentioned in Castelvechhi (2017) and Sturgis et al. (2018) and the 2016 US presidential election, mentioned in Zeedan (2019) and Breur (2016).

## 2.2 Political polls based on social media

Social media has become one of the most popular tools for exchanging information on the Web. By using social media sites such as Facebook or Twitter, people can share a lot of information and opinions. In this context, political polls based on social media have emerged as a cheap, accessible to a large portion of the population, and fast alternative to the traditional political polls. With social media, it is possible to infer opinions directly from the posts instead of asking users about their votes or political leaning. In this way, the opinions of a large sample of users can be gathered automatically and at a much faster pace. In the political context, social media is also being used for other purposes; for instance, it can be used to implement strategies to spread electoral campaign (Cornfield 2008).

The most popular ways for predicting elections based on social media rely on Taboada et al. (2011) and Xie et al. (2016): (i) volume and (ii) content. Volume-based approaches are the ones that only consider the volume of social media posts related to one candidate or political party in the sample collected from social media. Such approaches hypothesize that the higher the number of tweets mentioning

one candidate or political party, the higher the number of votes to the given candidate or party (Tumasjan et al. 2010). Some researches criticize this approach since it measures the political attention toward a political candidate or party and not the political support to a candidate or party (Jungherr et al. 2012), i.e., the volume mentioning a candidate or party is not necessarily correlated to a respective political support (Bovet et al. 2018). According to some lines of research such as Kassraie et al. (2017) and Srivastava et al. (2015), a candidate or party can be mentioned a lot in social media, but in a negative sense. On the other hand, content-based approaches (also called sentiment-based approaches) are the ones that use sentiment analysis and consider the meaning of the content within the collected sample. Basically, content-based approaches are based on the hypothesis that the higher the ratio of the number of positive posts related to a candidate or political party per the sum of the positive posts related to all candidates or political parties, the higher the chance of the given candidate or political party to be the winner of the election (Srivastava et al. 2015), by assuming that this ratio is proportional to the real vote.

### 2.3 Opinion mining and sentiment analysis

The idea of automatically inferring the sentiment polarity (positive, negative, neutral) of a sentence is not new among linguistic researches (Taboada 2016). This topic is popularly called *sentiment analysis* and is widely adopted in *opinion mining* tasks.

Techniques for sentiment analysis can be classified as follows (Taboada et al. 2011):

- *Lexical methods*: they are the methods that use dictionaries that relates a word to a sentiment (positive, negative, neutral). In this way, the calculus of the polarity of a sentence is based on the semantic orientation of the words that belong to it. Therefore, each word is associated with a (positive or negative) score, and the sum of word scores belonging to a sentence results in its final score. Usually, those methods can be associated with a set of predefined rules that can change the score of the words in a sentence when combined. For instance, when a negation term (“not”) precedes a word, its score can be discarded or considered as a negative one. An example of lexicon resource that can be used with this method is the Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al. 2001), which groups words according to different categories, including the *posemo* category, which stands for positive emotion, and *negsemo* category, which stands for negative emotion.
- *Machine learning methods*: these are methods that require the training of a model using previously labeled sentences (as positive, negative or neutral). They are

called supervised methods since they depend on labeled data sources for training the model. These methods use a set of features to distinguish sentences with different sentiments.

## 3 Approaches for elections outcomes prediction

In what follows, we present the four common approaches for predicting political elections outcomes as well as we briefly describe each work we found: counting based, political alignment, event detection and popularity-based approaches. The works that we could not categorize into these four approaches we describe in the last subsection, named “Other Works.”

### 3.1 Counting-based approach

In our literature review, we grouped twenty eight works that proposed to predict the election winner based on counting methodologies. While some of them assume that the number of instances mentioning a political candidate or party can be used to forecast elections (volume-based approach), others assume that the counting should rely on the number of *positive* posts related to them, for example (sentiment-based approach). In what follows, we describe these works that follow what we call Counting-Based Approach.

The research presented by Maldonado and Sierra (2015) focuses on predicting the 2012 Dominican Republic presidential election outcomes. The authors proposed a methodology to collect tweets, as follows: (i) only tweets containing candidates’ names and nicknames or parties’ names related to the two leading candidates were considered; (ii) only one tweet per user was stored taking into account that each person only votes one time; (iii) the polarity of the tweets was calculated based on the SAS Sentiment Analysis Software;<sup>1</sup> and (iv) tweets classified as positives were grouped by candidate. The number of voting intentions was calculated based on tweet counting. Tweets were collected into three “virtual polls” in different periods: from April 16 to May 7, 2012; from May 14 to May 18, 2012; and from 9 a.m. to 4 p.m. on Election Day. An average of 25751 tweets were collected in each “virtual poll.” Their results predicted the winner of the election correctly.

Almeida et al. (2015) proposed to predict the outcomes of six 2012 Brazilian municipal elections. They collected over one million tweets (to predict the six elections) posted by 171,680 users during the period of 10 weeks before the

<sup>1</sup> The link for the SAS Sentiment Analysis Software was not informed by the authors in Maldonado and Sierra (2015).

election. They discarded spam and news media accounts using an existing dataset of spammer/not-spammer users, and a set of news media accounts that were manually labeled. Users were characterized according to demographic characteristics based on their last tweets. In this step, they tested the multinomial Naive Bayes (MNB), SVM, and random forest (RF). To predict user gender, they searched for the name in a dictionary of names. If the name is not found, the classifier trained on user texts is invoked. Similarly, age is inferred based on a manually labeled dataset with texts of users of different ages. To predict social class, they adopted a classifier that was built filtering users by the places they visit (based on Foursquare data) and performing regional expression analysis. The MNB provided the best results. Finally, the sentiment of tweets mentioning only one of the candidates were predicted using: (i) a lexical dictionary; (ii) a MNB classifier trained with tweets labeled using emoticons; and (iii) a MNB classifier trained with a small set containing manually labeled tweets. They computed the votes of users to candidates testing two approaches: (i) each Twitter account represents one vote and is counted once, and (ii) different tweets of the same user are considered in the counting. The final sentiment was decided using majority voting. Their results were compared with the election official results and results of traditional polls that were performed by two main organization polls. The results related to two out of six cities were correct for the first round, and four out of six for the second round.

Dwi Prasetyo and Hauff (2015) tried to predict the outcomes of the 2014 Indonesia's presidential elections. They collected more than seven million tweets (from April 15 to July 8, 2014) containing the candidates' names or mentions to the candidate Twitter accounts. In order to determine the influence of non-personal and spam users in the data collected, they randomly selected a subset of users from the data collected and manually labeled them as either spam, non-personal (account of an organization) or "slacktivist" (newly user accounts whose content is totally related to elections). This labeling was based on the user-profiles and the most recent tweets. After that, they specified a set of rules based on the regularity of posts and terms posted to discard bot, slacktivists and nonpersonal accounts. In order to reduce bias, this approach weighted the contribution of tweets to votes depending on their geographical location and demographic information. The sentiment analysis was performed using a Naive Bayes (NB) classifier that was built from a dataset automatically labeled according to emoticons. The results of their most-basic predictor outperformed the majority of traditional polls, and their best performing predictor outperformed all the traditional polls.

Ibrahim et al. (2015) presented an approach to predict the 2014 Indonesian Presidential Election results. The Indonesian Sentiment Lexicon Vania et al. (2014) was applied to

the data in order to detect the polarity of the tweets collected. Ten million tweets containing the name or nickname of one of the two candidates were collected during the campaign period (from May 1, 2014 to July 6, 2014). After collecting tweets, the stop words, punctuation-mark, links, tweet mentions ("@"), and retweet characters (RT) were discarded. They have built a classifier to detect buzzer accounts, which are the ones that praise only one candidate and criticize/vilify the other ones. Buzzers usually are users with short-term creation date that produce a high frequency of tweets and retweets. The election prediction was calculated by leveraging the number of positive tweets for each candidate. Their results were better than predictions by several independent polling organizations.

Khatua et al. (2015) presented an approach to predict the results of the 2014 Indian General Elections. From March 15, 2014 to May 12, 2014, they collected four million tweets containing an abbreviation of political parties and candidates, and, to capture temporal events and sentiments, a set of dynamic keywords was extracted from the top-10 daily trending topics across the 15 politically sensitive cities in India. After collecting tweets, they removed URLs, duplicated tweets, and those mentioning more than one candidate/party or written in regional languages (only tweets written in English were considered). They observed that the few tweets with location information were uniformly distributed along India, i.e., it had no location bias. A word frequency list was generated considering the entire dataset, and the most relevant and contextual keywords were manually selected in order to filter for relevant data. The regression technique ordinary least squares (OLS) was applied for predicting vote and seat share. The sentiment analysis was performed based on two lexicons. This research concluded that the volume method, as well as the sentiment analysis method, predicted the results of the elections correctly.

Srivastava et al. (2015) proposed to predict the outcomes of the 2015 Delhi Assembly Election building a training dataset that combines: (i) the IMDb, a dataset that contains movie reviews classified as positive or negative; and, from the total of 3,052,730 tweets collected during about 1 month related to Delhi elections: (ii) a subset manually annotated (as positive or negative); and, similarly, (iii) another subset automatically labeled based on emoticons. Tweets that contain positive *and* negative emoticons were discarded, i.e., only tweets associated with emoticons of one polarity were considered. To detect whether the tweet expresses a sentiment related to a candidate, a list of positive and negative emoticons and a word list from the SENTIWORDNET (Esuli and Sebastiani 2007) was adopted. The sentiment analysis was performed using SVM with unigrams and bigrams. The ratio of the number of positive tweets related to a party with the sum of the positive tweets related to each party was calculated (sentiment share). The seat share was

calculated, in turn, based on the sentiment share. Although the number of seats per party obtained with this method was not the same as the real numbers, the overall results were closer to the real ones, and the order of the parties with the most seats was correctly predicted.

Burnap et al. (2016) proposed a method to forecast the 2015 UK General Elections. For about 4 months, more than ten million tweets were collected (containing political parties' names and candidates' names). They analyzed the sentiment of the tweets using a lexical approach. The scores of all tweets with positive sentiment were summed in order to obtain the overall magnitude of the positive sentiment and avoid a tie between candidates that have the same number of positive tweets. Tweets containing the name of more than one candidate/party were discarded to avoid misallocating the positivity of the tweet. The positive sentiment scores of each party were combined to calculate the total sentiment, which is used to normalize the positive party sentiment sum for each party concerning the other ones. Although their numerical results did not reflect the real ones, this approach predicted the order of the first three parties correctly.

Jose and Chooralil (2016) presented an approach to predict the results of the 2015 Delhi elections. The prediction method combines different machine learning and lexicon-based classifiers, namely Naive Bayes, Hidden Markov Model, and SentiWordNet (SWN), to reduce the risk of selecting an inappropriate classifier. The SWN assigns different sentiment scores to different words, depending on the role of the word is used in the sentence ("part of speech"). The majority voting rule was used to classify a tweet as positive or negative when the classifiers return different responses. In addition, a negation handling method was adopted. About 12,000 tweets were collected for 3 weeks during Delhi elections, by searching for the names of two candidates. With their analysis, the authors concluded that the positive sentiment toward one of the candidates was clearly higher than that of the other one. However, they did not inform if such a result was the same as the real election outcomes.

Sharma and Moh (2016) described an approach to predict the outcomes of the 2016 Indian general state elections. Since it would be challenging to predict the results considering different dialects spoken in that country, they used a specific tool to capture only tweets written in the Hindi language. A total of 42,235 tweets were collected during 1 month, and only tweets containing party names were considered. During preprocessing, URLs, hashtags, Twitter mentions, stop words, emoticons, special characters, and punctuations were removed. A negation handling method was adopted to identify words of Hindi language that can revert the meaning of a sentence (as occurs with the words *no* and *not* in the English language). A dictionary-based approach was adopted to automatically assign tweets

polarities (positive, negative, or neutral). After that, they conducted another experiment in which 36,465 tweets were manually annotated. Three different methods were adopted to sentiment analysis, namely SVM, NB, and a dictionary-based approach. While the prediction obtained with the dictionary-based approach did not reach the correct result, the final result calculated with the SVM and NB classifiers predicted the winner party of the 2016 Indian general state election correctly.

Wicaksono et al. (2016) proposed a method to predict the results of the 2016 US presidential election. They used the Binary Multinomial Naive Bayes and Sentiment140 tweet corpus to classify tweets. A total of 400 tweets whose localization field is one of the 51 electoral colleges and whose text contains election keywords, party, or candidate names were analyzed. They were collected from August 7 to August 15, 2016. Abbreviations and contractions were expanded with the aid of a dictionary, aiming at improving the analysis. During the preprocessing step, HTML, URL, mentions, non-alphanumeric, and stop words were removed from the data collected. A tweet with positive sentiment is considered a vote to the mentioned party/candidate. On the other hand, if a tweet has negative sentiment, it is considered a vote to the opposite of the mentioned party or candidate. The winner in a state is the one who received the majority of the votes. In order to implement the electoral vote, the electors who win in its state must cast their votes to the party that appointed them. Finally, the party/candidate who got the most electoral votes is predicted as the election winner. Although from the 51 electoral states, only 48 were associated with Twitter data, their results predicted the election winner correctly.

Sanders et al. (2016) presented a study to predict two Dutch elections using Twitter, namely the 2012 national parliamentary elections and the 2015 provincial elections. A total of 159,826 political tweets was collected for the 2012 election, and a total of 183,602 tweets was collected for the 2015 election. The prediction method consists in counting how often a political party is mentioned in the set of tweets posted in the 10 days before and including Election Day. A second prediction method was tested by considering demographic information (age and gender) automatically inferred from the history of Twitter users posts using the TweetGenie (Nguyen et al. 2014) software. Adding demographic information only improve the prediction for the 2015 provincial elections. This approach was able to predict correctly the election winner.

Vepsäläinen et al. (2017) proposed to predict the results of the 2015 Finnish parliamentary elections based on the number of likes gathered from candidates' official pages on Facebook. About 2.7 million Facebook likes were gathered from candidates' official pages on Facebook. The data collection occurred four times within 1 month before the election. A statistical forecast model that considers each Like on



a candidate's page as one vote was adopted to predict results. It was observed that Facebook users from Finland do not represent the Finnish electorate after analyzing demographics information of the users. The authors observed that there is a positive relationship between votes and Facebook likes but concluded that election prediction based on Facebook likes was less accurate than the results obtained by traditional polls.

Heredia et al. (2017) tried to predict the 2016 US general elections based on data from Twitter considering two approaches: volume-based and sentiment-based, focusing on the two most popular candidates. The collected data were labeled using a convolutional neural network that was trained using the sentiment140 dataset, whose tweets were automatically labeled using emoticons. In this way, the final training dataset was composed of 1.6 million tweets (800,000 positive and 800,000 negative). Three million tweets containing terms related to the two candidates were collected from September 22 to November 8 in order to be classified using the trained network. During the preprocessing step, emoticons, retweets, and duplicated tweets were removed. The election dataset was separated into seven different sets according to different periods. The results were compared with three traditional polls conducted during 13 days before the election. The authors concluded that in their experiments, the volume-based approach was very inaccurate. On the other hand, the sentiment-based approach presented results inline with the results of the most accurate poll in previous elections.

Rosseti et al. (2017) presented an approach that collects tweets to predict the results of the 2016 US presidential election. During the presidential election campaign, a total of 1,974,401 posts (430,529 tweets and 1,543,872 retweets) were collected from October 6 to November 7, 2016. Those tweets were posted by 432,289 users, an average of 4,56 posts per user. In addition to collect the text of the tweet, they collected the number of likes of each tweet, the username, description, and location of the user. Tweets were collected by filtering by the ones that contain candidates' names and restricting by geolocation parameters (latitude, longitude, and ratio) related to the USA. Although the authors collected user profile information, they did not use it to do the election forecast. The prediction was made using the volumetric approach, which consists in counting the number of tweets mentioning each candidate. The real winner of the 2016 US election, namely Donald Trump, was mentioned only in 3% of the total tweets collected. Then, their results did not correctly predict the winner of the election.

Ramzan et al. (2017) have tried to predict the winner of the 2017 Indian UP state election using sentiment analysis. A total of 10,000 tweets were collected in 1 week. In the preprocessing step, tweets were converted to lowercase, and hashtags, extra spaces, some stop words, URLs, and usernames have been removed. The algorithm used to perform

the sentiment analysis, and the keywords used to collect data were not informed. The distribution percentage of tweets has been calculated to each party, and the one that received the highest fraction of positive sentiments is considered the election winner. The experimental results were compared with the actual results and predicted the actual winner correctly.

Hinch (2017) presented a research that proposed to predict the 2016 US Presidential Election using Twitter. A total of 8696 tweets for Clinton and 4004 tweets for Trump were collected from October 2015 through November 2016. Only tweets geotagged to the states of Wisconsin and Michigan were considered. Collection keywords were related to candidate names and campaign slogans. The prediction method consists in counting the number of tweets mentioning each candidate in each state. While the tweets related to Michigan and Wisconsin voters were aligned with traditional polling methods, neither technique successfully predicted a Trump win in both states.

Singh et al. (2017) presented an approach to predict the 2016 Spanish General Elections using Twitter. They collected a total of 90,154 tweets from June 6 to June 26, 2016 using hashtags related to the party and candidate names. The preprocessing step removed hashtags and web links. They built their own lexical dictionary and considered that a tweet is positive when it contains a positive word. After computing the number of positive tweets related to each one of the political parties, they concluded that although results of second and third place party were predicted incorrectly, this approach was able to predict correctly the winner and the political party in fourth place.

Praciano et al. (2018) adopted a sentiment analysis counting approach and Twitter data to predict the outcomes of the second round of the 2014 Brazilian presidential elections. A total of 158,279 were collected during October 12, 2014 and October 28, 2014, considering tweets mentioning candidates' names. Besides the tweet text, other information such as author's name, date, count of retweets and localization also were stored. Sentiment analysis was conducted using lexical dictionaries, namely TextBlob (translating the tweet text for English to be able to get the sentence polarity using TextBlob) and OpLexicon combined with Sentilex. The preprocessing steps removed tags, profile mentions, links, punctuation, Portuguese stop words, unnecessary spaces and converted words to lower case. Support vector machine (SVM), NB, logistic regression (LR) and decision trees (DT) were used in their experiments. A space-temporal analysis was also conducted to find out the general opinion by state and over time. The best result (considering accuracy, precision, recall and F1 score) was achieved with the SVM algorithm, where for many states (but not all), this approach pointed out the candidate who received more votes correctly, comparing their outcomes to data from the superior electoral court database.

Heredia et al. (2018) tries to predict the outcomes of the 2016 US general elections on national and state levels. The prediction was performed by calculating tweet volume and positive sentiment per candidate (considering the two front-runners candidates). They collected about three million tweets ranging from September 22 to November 8, 2016, using terms related to elections or candidates' names. For state-level analysis, tweets are selected based on the location information found in the user profile. The data were normalized based on the number of electors in each state. A deep neural network (AlexNet) was trained using sentiment140 dataset, where emoticons were used to label tweets as positive or negative. Since neural networks require numerical inputs, they used character embeddings. The voting ratio was adjusted by dividing the count of votes for one candidate by the total votes for both candidates. Their results were compared to different opinion polls. They concluded that, on the national level, volume trends are similar to trends in positive sentiment, and sentiment analysis of election tweets provides values that are close to the traditional election polls. On the other hand, at the state level, the volume is not a useful metric for predicting elections, and the results obtained with sentiment analysis did not match the elections results for some states.

Bilal et al. (2018) proposed to predict the outcomes of the 2018 Pakistan General Elections. In order to do that, they trained a recurrent neural network (RNN) using tweets related to the 2013 elections. It is not described how tweets related to the 2013 general elections were gathered and how was the labeling method that was adopted. A total of 65,000 tweets were collected from the mid of May 2018 by the mid of July 2018, verifying tweets containing top trending hashtags and tweets mentioning party/leader names. Approximately 55,650 tweets related to the 2013 General elections were collected from a relevant research. The preprocessing steps removed blank spaces, double spaces, usernames, the word "RT" (retweet), emojis, links, and punctuation. For converting the textual data into numeric representation (tokenization) and for making each sentence of the same length (padding), they used Keras.<sup>2</sup> The validation set contained 10% tweets from each dataset (2013 and 2018). One of the parties of the 2018 dataset did not exist in 2013. Therefore, this party was only included in test dataset as unseen data. The strategy predicted the winner party correctly, achieving results close to the real outcomes.

Naiknaware and Kawathekar (2018) proposed to use a sentiment analysis method to predict the 2019 India election. Tweets were collected based on hashtags related to the election, and the polarity was manually assigned. The exactly collection period was not informed. The preprocessing

step converted the text to lowercase, removed hashtags, stop words, extra spaces, the term "RT" (retweet), punctuations, URLs, usernames, and replaced repeated characters with just two of that letter. A slang dictionary was used to replace slangs by their associated meanings. The sentiment score of each tweet is computed, as follows:  $score = sum(pos.matches) - sum(neg.matches)$ . The authors argue that the 2019 main elections agendas are probably GST, Demonetization, Digital India, Make in India, Startup India, Swachh Bharat, Kashmir and Yoga day. The idea of them is to apply a counting strategy to these agendas using sentiment analysis and to determine which agenda is most supported by the population. By analyzing tweets, they verified that Demonetization and Kashmir are associated with negative opinions and GST, Digital India, Make in India, Startup India, Swachh Bharat and Yoga day are associated with positive opinions and concluded that in 2019 same government is to be elected.

Bansal and Srivastava (2018) proposed to predict the results of the 2017 Uttar Pradesh (U.P) legislative elections using Twitter. They collected more than 300,000 tweets from February 1 to 20, 2017. The keywords adopted to search data were as follows: party names, party leader names, multiple official election campaign handles. Geo tagging was employed in the cases where keywords are not exclusive to the elections. Preprocessing included converting into lowercase, stemming, and removal of white spaces, punctuations, symbols, numbers and stop words. They find out tweet topics using word co-occurrences, with both Latent Dirichlet Allocation (LDA) algorithm and Bitern Topic Model<sup>3</sup> (BTM). The sentiment score and polarity of a tweet are labelled according to the assumption that a tweet is a mixture of weighted topics, whose sentiment is assigned based on the Sentiwordnet. Election results were computed using three methods: (i) computing total volume of tweets per party; (ii) computing total volume of positive tweets per party; and (iii) computing total positive magnitude of tweets per party. Methods (ii) and (iii) were able to predict the correct position of the main parties of the given election and (ii) achieved the closest result in relation to the real outcomes according to vote share.

Budiharto and Meiliana (2018) presented an approach to predict the 2018 Indonesian Presidential elections using Twitter. Tweets from two political candidates of the given election were collected from March to July 2018. The average number of likes and retweets of the candidate posts was calculated for each one of the two candidates. Also, tweets with hashtags related to election hashtags were gathered for sentiment analysis. Preprocessing steps included removal of URLs, stop words, special characters and duplicated tweets.

<sup>2</sup> <https://keras.io/>.

<sup>3</sup> <https://github.com/xiaohuiyan/BTM>.

The sentiment polarity of tweets was calculated using TextBlob. The candidate associated with the highest number of retweets and likes is also associated with the higher number of positive tweets. By comparing this analysis with the real election outcomes, this approach was able to predict the election winner.

Bansal and Srivastava (2019) proposed to predict the results of the 2017 Uttar Pradesh (UP) legislative elections using Twitter. They collected 300,000 tweet during 8 days by searching for party and candidate names and keywords related to campaign slogans. Keywords that are not exclusive of this election (e.g., congress) were restricted to geo-location of UP. Preprocessing steps included: convert text to lower case, removal of stop words, white spaces punctuation, numbers, and duplicated tweets. To obtain vote share prediction to each party, they summed all positive sentiments for each party and normalized it by using the total positive sentiment. The sentiment of tweets were inferred using lexical dictionaries. Also, they adopted an emoji lexical dictionary to automatically label tweets according to emojis. Election prediction was computed based on: (i) computing total positive magnitude by party and (ii) counting positive tweets by party. This research was able to predict correctly the position of the political parties in all the experiments. Results obtained by considering emojis polarity were better than the ones that only consider words polarity. Furthermore, results computed using only tweets polarity were better than the ones that consider tweets positive magnitude.

Hwang (2019) proposed to predict the outcome of the 2016 US presidential elections using data from Reddit, a popular online forum. This approach gathers posts and responses to posts that contain any (case insensitive) combination of the candidates names, restricting by 8 dates/periods that correspond to political key dates/periods, namely July 18–21, 2016; July 25–28, 2016; September 9, 2016; September 26, 2016; October 7, 2016; October 9, 2016; October 28, 2016; November 7, 2016. After that, they performed a volume-based approach that counts how many times each candidate was mentioned in the posts. Also, they performed a sentiment-based approach using Aylien API<sup>4</sup> to determine if Reddit posts mentioning the political candidates are positive, negative or neutral. In the sentiment-based approach, neutral posts are discarded and the probability of a candidate to win the election is given by the following formula:  $p(c_1) = \frac{pos(c_1)+neg(c_1)}{pos(c_1)+neg(c_1)+pos(c_2)+neg(c_2)}$ . The volume-based approach was not enough to predict the election winner correctly. This research shows that the sentiment of relevant Reddit posts reflected the results from election polls.

Kristiyanti et al. (2019) presented a method to predict the outcomes of the 2019 presidential elections of the Republic of Indonesia using Twitter. They adopted the SVM algorithm with selection features of particle swarm optimization (PSO), bi-grams and genetic algorithms (GA). A total of 4000 tweets was gathered using keywords related to the Indonesian election and candidates names during the campaign period until the presidential election in April 2019. The exactly collection period was not informed, and the tweets were labeling according to hashtags. The preprocessing step removed punctuations and special characters. To predict the outcomes, they use a sentiment-based counting strategy. This method was not able to predict the election winner correctly.

Singh et al. (2020) tried to predict the outcomes of the 2017 Punjab (a state of India) assembly elections using Twitter. They use sentiment analysis to predict the number of seats that the political parties are likely to win in the election. A total of a total of 9157 tweets were collected over a period of 28 days based on hashtags related to the political parties and candidates names (for three parties). In addition to tweets written in English, tweets written in Punjabi (local language) were also collected. The translation from Punjabi to English was performed by an expert team. Preprocessing steps converted all words to lower case and removed extra blank spaces, English stop words, links, punctuations and numeric values. The sentiment analysis step was divided into: (i) emotion analysis (trust, surprise, sadness, joy, fear, disgust, anticipation and anger)—using the syuzhet package in R-language); (ii) polarity analysis—they manually annotate 1000 tweets (500 positive and 500 negative), annotated datasets of two other domains (amazon and IMDb reviews). The algorithms used were as follows: DT, K-Nearest Neighbor (KNN), and SVM. According to the authors, SVM outperformed the other models to compute the sentiment polarity. The seat forecast method uses the polarity analysis and historical data (i.e., results of previous elections—election year, vote share, number of seats, winner political party). First of all, the actual sentiment score (ASS) is computed, a s f o l l o w s :  $ASS = \frac{(PSS \text{ of party } P) - (NSS \text{ of party } P)}{\sum (PSS \text{ for all parties}) - \sum (NSS \text{ for all parties})} \times 100$  where PSS refers to the total number of positive tweets for a party P, and NSS to the total number of negative tweets for a party P. Next, linear regression was used to the seats forecasting, using ASS information and considering the vote share as independent variable and the number of seats as the dependent variable. The number of seats predicted were not equal but were very close to the real ones, predicting the winner party correctly.

Sanders and van den Bosch (2020) adopted Twitter as source of opinions and a counting strategy to predict the Dutch elections of the parliament in 2012 and for

<sup>4</sup> <https://docs.aylien.com/textapi/endpoints/#api-endpoints>.

**Table 2** Counting-based approach summary

Paper	Year	Type	Country	Vol.	Sent.	ML	Sample	Success
Maldonado and Sierra (2015)	2012	Presidential	Dominican Republic	–	✓	N/A	25,751	✓
Almeida et al. (2015)	2012	Municipal	Brazil	–	✓	✓	1,000,000	Partial
Dwi Prasetyo and Hauff (2015)	2014	Presidential	Indonesia	–	✓	✓	7,000,000	✓
Ibrahim et al. (2015)	2014	Presidential	Indonesia	–	✓	✓	10,000,000	✓
Khatua et al. (2015)	2014	General	India	–	✓	✓	4,000,000	✓
Srivastava et al. (2015)	2015	State	India	–	✓	✓	3,052,730	✓
Burnap et al. (2016)	2015	General	United Kingdom	✓	–	–	10,000,000	✓
Jose and Chooralil (2016)	2015	State	India	–	✓	✓	12,000	N/A
Sharma and Moh (2016)	2016	State	India	–	✓	✓	42,235	Partial
Wicaksono et al. (2016)	2016	Presidential	USA	–	✓	✓	N/A	✓
Sanders et al. (2016)	2012, 2015	Parliamentary, provincial	Netherlands	✓	–	N/A	159,826, 183,602	✓
Vepsäläinen et al. (2017)	2015	Parliamentary	Finland	✓	–	–	2,700,000	N/A
Heredia et al. (2017)	2016	General	USA	✓	✓	✓	1,600,000	Partial
Rosseti et al. (2017)	2016	Presidential	USA	✓	–	–	1,974,401	No
Ramzan et al. (2017)	2017	State	India	–	✓	N/A	10,000	✓
Hinch (2017)	2016	Presidential	USA	✓	–	–	12,700	No
Singh et al. (2017)	2016	General	Spain	–	✓	–	90,154	✓
Praciano et al. (2018)	2014	Presidential	Brazil	–	✓	✓	158,279	Partial
Heredia et al. (2018)	2016	General	USA	✓	✓	✓	3,000,000	Partial
Bilal et al. (2018)	2018	General	Pakistan	–	✓	✓	65,000	✓
Naiknaware and Kawathekar (2018)	2019	General	India	–	✓	N/A	N/A	N/A
Bansal and Srivastava (2018)	2017	Legislative	India	✓	✓	✓	300,000	Partial
Budiharto and Meiliana (2018)	2018	Presidential	Indonesia	✓	✓	–	N/A	✓
Bansal and Srivastava (2019)	2017	Legislative	India	✓	✓	–	300,000	✓
Hwang (2019)	2016	Presidential	USA	✓	✓	N/A	N/A	Partial
Kristiyanti et al. (2019)	2019	Presidential	Indonesia	–	✓	✓	4000	No
Singh et al. (2020)	2017	State	India	–	✓	✓	9157	✓
Sanders and van den Bosch (2020)	2012, 2011, 2015	Parliamentary, provincial, provincial	Netherlands	✓	✓	–	N/A	✓

the provinces (and the senate) in 2011 and 2015. Tweets containing names of political parties were collected during 10 days before elections. A total of 17,000 tweets were annotated by at least three annotators in regard to some features denoting communicative intent (such as the presence of sarcasm, sentiment polarity, the presence of an explicit voting advice or voting endorsement, etc.). These annotations were used to create filters for exclusion of tweets with certain combinations of features, for example, for removing all sarcastic tweets, or for removing all tweets in which the person posting the tweet explicitly states that he or she will not vote for a particular party. Grid search was applied over all possible filters to compute the lowest MAE prediction error for each one of the three elections. They concluded

that the filters achieved different behavior for the different elections and only a small MAE improvement is possible when optimizing on all three elections.

### 3.1.1 Summary

The general characteristics of the works presented in this section are summarized in Table 2. The column *paper* refers to the paper, the column *year* refers to the year of the election that is being predicted in the paper, column *type* indicates the type of the election; the column *country* refers to the election country; column *vol.* is checked if the paper presents a volume-based approach; the column *sent.* is checked if the paper presents a sentiment-based approach; the column

*ML* is checked if the paper uses machine learning; the column *sample* refers to the quantity of data collected; the column *success* indicates if the approach achieved success to predict the correct election winner. The term *N/A* is used to indicate that the paper does not present enough information to fill the table field.

In Table 2, we can observe that 17 out of 28 works analyze presidential or general elections. Although these scenarios involve too many electors, the sample size varies a lot. We can also observe that 15 out of 28 works use machine learning algorithms, and 22 use sentiment analysis techniques. The total of 15 out of 28 works succeeded in predicting the election winner correctly.

The works in Almeida et al. (2015), Praciano et al. (2018), and Heredia et al. (2018) are classified as partial success since their method were successful only for some cities/states. Similarly, the works in Heredia et al. (2017), Bansal and Srivastava (2018), and Hwang (2019) are classified as partial success as they presented both volume-based and sentiment-based methods, achieving success for the sentiment-based method and failing when using the volume-based method. Finally, the work in Sharma and Moh (2016) is classified as partial success as it achieved success using sentiment analysis with machine learning and failed using sentiment analysis based on sentiment dictionaries.

Considering that this type of approach has been adopted by many works, as follows we present a brief discussion about works that use machine learning and those that do not, and works that use sentiment analysis and those based only on volume.

### 3.1.2 ML versus no ML

By comparing counting-based approaches that use machine learning methods and the ones that do not use, we observed that 10 (66.67%) out of 15 approaches that adopted ML succeeded (including the cases Bansal and Srivastava 2018; Heredia et al. 2017; Sharma and Moh 2016)—listed in Table 2 as partial success as they adopted more than one prediction method but predicted the correct election winner when using ML techniques). Only 1 paper (6.67%) out of 15 (100%) failed completely to predict the election winner. On the other hand, 5 (62.50%) out of 8 papers that do not use machine learning algorithms succeed and 2 (25%) out of 8 failed completely to predict the election winner. These findings suggest that the use of machine learning methods may improve predictions of counting-based approaches.

### 3.1.3 Volume versus sentiment

We compared the volume-based counting approaches with the sentiment-based counting approaches as follows. From the total of 23 papers that presented methods based

on sentiment, 2 (Vepsäläinen et al. 2017; Naiknaware and Kawathekar 2018) cannot be analyzed since their success is not informed. The works in Heredia et al. (2017), Bansal and Srivastava (2018), and Hwang (2019) presented both volume-based and sentiment-based methods and obtained success only in the sentiment-based case. Also, from the total of 12 papers based on volume, 1 (Vepsäläinen et al. 2017) is not included in this analysis since its success is not informed. We observed that 5 (45.45%) out of the 11 remaining (100%) volume-based approaches succeeded. On the other hand, 16 (76.19) out of 21 (100%) papers that presented a sentiment-based counting approach were successful (including Heredia et al. 2017; Bansal and Srivastava 2018), and (Hwang 2019) that presented both volume-based and sentiment-based methods and achieved success using the sentiment-based methods and failed when using the volume-based method). These findings suggest that sentiment may improve predictions of counting-based approaches.

## 3.2 Political alignment approach

We grouped five works that try to forecast elections results by predicting the political leaning/alignment of the users. In this section, we present a brief summary of these works that follow what we call Political Alignment Approach.

Bachhuber et al. (2016) proposed to predict the results of the 2016 US Presidential Elections using Twitter. They manually identified approximately 50 supporters per candidate and used their tool called Voter-Profile to analyze the frequency of basic word groups in the supporter posts: articles, negations, auxiliary verbs, conjunctions, prepositions, pronouns, quantifiers. The frequencies of word groups were used as training data for the Decision Tree algorithm. After observing that the linguistic differences between supporters of different candidates were too small to make reliable predictions, they ended up by changing the classification task to distinguish only between two classes, namely *democrats* and *republican* users. However, even in this case the linguistic analysis was not able to predict election results correctly.

Castro and Vaca (2017) tried to predict the citizens' political alignment in order to predict the result of the 2015 Venezuelan Parliamentary Elections. Approximately 750,000 tweets (from October to December 2015) were collected and organized in three different datasets: (i) posts produced by people from Venezuela (using geolocation information); (ii) Twitter data from ten government political leaders; and (iii) data from the Twitter of ten opposition leaders. The preprocessing of tweets included stemming and removal of stop words, punctuation, and emoticons. Also, bots accounts have been discarded. They considered that bots are accounts with many friends and a few followers and calculate a reputation score to categorize whether an account is a bot. A dictionary was built containing the most common political

**Table 3** Political alignment approach summary

Paper	Year	Type	Country	Alignments	Sample	Success
Bachhuber et al. (2016)	2016	Presidential	USA	Democrats republicans	N/A	No
Castro and Vaca (2017)	2015	Parliamentary	Venezuela	Government opposition	750,000	✓
Bastos and Mercea (2018)	2016	Referendum	United Kingdom	Populism economism nationalism globalism	1,295,956	No
Campanale and Caldarola (2018)	2016	Referendum	Italia	Yes-oriented no-oriented uncertain	1,295,956	N/A
Lopardo and Brambilla (2018)	2018	Midterm	USA	Democrats republicans	160,000	Partial

terms considering the content posted by the political leaders of the two lines (government and opposition). These terms were extracted by using the LDA algorithm. Basically, if a user posts a tweet with any of such terms, his political leaning can be inferred. A random set of Twitter users that posted political content was selected and manually labeled by independent annotators as “Government,” “Opposition” or “Ambiguous.” SVM was adopted to classify users as “Government” or “Opposition.” The result obtained with this approach reflected the election results.

The work described by Bastos and Mercea (2018) presented a study that uses data from social media to identify political/ideological alignments about the Brexit debate and determine its result. They use geolocation information and a deep learning algorithm in order to detect a division about globalist and nationalist standpoints. The idea of that approach was to calculate the ideological leaning (populism, economism, nationalism, globalism) of the users and map them according to voting constituencies (England, Wales, Scotland, and Northern Ireland). Then, the authors model the prevailing ideological opinion of each parliamentary constituency. They collected about eight million tweets for approximately 2 months and relied on two expert coders who classified a subset of tweets along with the ideological coordinates. The deep learning algorithm identifies at least one and a maximum of two ideological coordinates because globalism-nationalism and economism-populism are mutually exclusive ideological alignments. The classifier identified a robust nationalist sentiment throughout the campaign and an almost equal division between nationalist and globalist sentiments in the last days of the campaign. The model presented was not able to predict the outcome of the referendum.

Campanale and Caldarola (2018) proposed to predict the outcomes of the 2016 Italian Constitutional Referendum using Twitter. A total of 1,200,000 tweets were collected using the tag “referendum” during the month preceding the polling day (from November 1 to December 3, 2016) to create the dataset. They collected exactly 1,295,956 tweets. The political orientation (yes-oriented, no-oriented, uncertain) of the tweets is predicted according to their hashtags (the three most representative hashtags by class were selected according to the “Rite Tag” website) During text

preprocessing, truncated words, URLs, Italian stop words, snails and hashtags were removed. They built a MNB classifier. For each user, his class was also computed by taking into account the number of tweets related to the “yes”/“no” classes. Whenever an “uncertain” is identified, they verify if the user who posted it belongs to the class “yes” or “no.” If the user belongs to the class “yes”/“no,” then his tweet previously classified as “uncertain” is changed to be “yes”/“nos.” On the other hand, if the user class is not defined, then the tweet class is “uncertain.” The authors state that they achieved promising results.

Lopardo and Brambilla (2018) proposed to use Twitter to predict the support to two major US political parties in the 68 most competitive House of Representative districts during the 2018 mid-term elections. The main idea was to predict the political alignment (democrats, republicans) of the users and based on that predict election results. To build the training and test datasets, they collected tweets of users with clear political alignment/affiliation such as political activists, candidates, partisan pundits, and organizations. Approximately 160,000 tweets posted in 6 months before Election Day were collected. Another experiment was conducted by adding 120,000 other tweets from less known users. They gathered tweets mentioning a national leader or party, posted from a limited region (posted from locations within each district). They adopted a RNN-Long short-term memory (LSTM) binary classifier to classify each tweet as being Republican or Democratic. Their method predicted the correct winner on 60% of the districts.

### 3.2.1 Summary

The general characteristics of the approaches presented in this section are summarized in Table 3. The column *paper* indicates the paper that is being analyzed, the column *year* refers to the election year, column *type* indicates the type of the election; the column *country* refers to the election country; the column *alignments* refers to the alignments that are being predicted; the column *sample* refers to the quantity of data collected; finally, the column *success* indicates if the approach achieved success to predict the correct election winner. The term *N/A* is used to indicate that the paper does not present enough information to fill the table field. In this

table, we can observe that this approach was not popular to predict presidential or general elections, as it failed in the only attempt to deal with this type of election (Bachhuber et al. 2016). Also, only one succeeded in predicting the election result. We classified the success of the work in Lopardo and Brambilla (2018) as “partial” because it predicted the correct results only for some districts.

### 3.3 Event detection approach

In this subsection, we group three works that relate the victory of a political candidate/party to the occurrence of political events and use social media to detect events occurrence in order to predict the election winner.

Unankard et al. (2014) proposed a method to predict the 2013 Australian Federal Election election outcomes at state and national levels based on sentiment analysis and subevent detection. First, tweets containing terms and hashtags related to this election are collected. After that, a preprocessing step is performed to (i) remove stop words, web addresses, retweet keyword, and username mentions; (ii) replace slang and word extensions like “boored” by English words; (iii) stem all words. In order to understand opinion in particular areas, they extract geolocation information from the tweets or user location from the user profile. When both data are not available, the user location is set to be “Australia.” After that, tweets related to the same event/topic are clustered together based on their terms. The cosine similarity function was used to calculate the similarity between an existing cluster and the new tweet, and every tweet is compared with all previous clusters’ centroids. A new cluster is created if the similarity is below to a given threshold to all existing clusters. The clusters created are considered subevents if there is a strong correlation between the event location (mentioned in the messages) and the user location. The event location is identified using a part-of-speech (POS) tagging to detect proper nouns, and a Named Entity Recognition technique together with the Stanford Named Entity Recognizer to detect locations. The sentiment analysis was conducted by using a lexicon dictionary that was expanded to deal with informal text and emoticons. In order to do that, they downloaded an existing slang list and manually annotated it. A POS tagging strategy was adopted, and adjectives, adverbs, verbs, nouns, interjections, emoticons, and hashtags were related to an opinion score. Rules to change the predefined score to each word are used in negation and intensification cases. Additionally, they propose some rules to try to identify sarcasm constructions. A total of 808,661 tweets were collected since the announce Election Day (August 4, 2013) until the day before Election Day (September 6, 2013), and only the two main parties were considered. User accounts whose username contains the words “news” and “TV” are removed because they can be media account. A sub-event

score, representing the significance of each sub-event topic, is calculated based on the number of tweets they are associated with. The voter preference is defined as the highest score out of the two candidates after summing up the scores of messages of each candidate. They downloaded a predefined list of events that took place in the same period, and a subset of tweets related to the Australia election was manually labeled by three annotators. Their results were compared with the actual results and opinion polls’ results and predicted the election results correctly.

Tung et al. (2016) presented a method to predict the outcomes of the 2014 Taiwan Mayor Elections. They analyzed comments related to political events and derive a set of rules from them. A rule can state, for instance, that if an event that has a positive/negative impact on a given candidate occurs days before the election, then he will win/lose the election. They assume that an event arises when the number of relevant words of the event is higher than usual and when there is a sudden surge of its popularity. For this reason, they calculate the average number of articles or comments related to a topic at different time points. A total of 155,921 published articles were collected from August 1 to November 28, 2014 in the PTT Bulletin Board System (BSS), which is the most popular BSS in Taiwan. Additionally, 5,532,824 comments related to these articles from 106,551 users were also considered in their analysis. The articles were preprocessed by using a word segmentation system to remove punctuation and stop words. A set of rules using a politician dictionary was used to decide which candidate the article is talking about, and the number of like/dislike tags per post was used to compute the support score to the given candidate. Comments were not analyzed. The event type is classified based on its influence (positive, negative, or useless) and size (small or big). Finally, the prediction model created the rules that will be used to predict election outcomes. They use the LDA model to identify relevant political topics in articles and detect events based on the number of papers or number of comments. Four proposed methods (PGP-A, PGP-C, SGP-A, and SGP-C) are analyzed to create event sequences. The methods PGP-A and PGP-C identify the subject of the article based on a grammar parser proposed by the authors. On the other hand, the methods SGP-A and SGP-C identify the subject of the article based on the Stanford Grammar Parser. The better results were obtained by adopting the methods PGP-A and PGP-C. Taking into account seven cities, they predicted the results of six cities correctly.

Shaban et al. (2017) proposed to predict the results of the 2016 US Presidential elections. They state that most of the events that occur during a campaign did not change the course of the election and investigate which events could, indeed, significantly affect and change the dynamics of the election. For this purpose, the authors collected about 135.5 million tweets over 6 weeks before the election (containing

**Table 4** Event detection approach summary

Paper	Year	Type	Country	Sample	Success
Unankard et al. (2014)	2013	General	Australia	808,661	✓
Tung et al. (2016)	2014	Municipal	Taiwan	5,532,824	Partial
Shaban et al. (2017)	2016	Presidential	USA	135,500,000	N/A

candidates' names or election terms) and political articles. News articles related to the elections were clustered (using vector space models), and keywords were extracted from them. Event keywords were also extracted from the tweets collected by using a method based on TF-IDF. Tweets related to more than one candidate were discarded. The analysis of tweets is performed considering tweets count, sentiment analysis, and win/loss terms (counting how many tweets have words similar to "win" and "lose," according to the WordNet). The sentiment analysis step was performed using a convolutional neural network with lexicon embeddings. The authors concluded that real events have a significant impact on elections and can be used to predict election results.

### 3.3.1 Summary

The general characteristics of the approaches presented in this section are summarized in Table 4. The column *paper* indicates the paper that is being analyzed, the column *year* refers to the election year, column *type* indicates the type of the election; the column *country* refers to the election country; the column *sample* refers to the quantity of data collected; finally, the column *success* indicates if the approach achieved success to predict the correct election winner. The term *N/A* is used to indicate that the paper does not present enough information to fill the table field. We can observe that this approach succeeded to predict the results of general elections using thousands of opinions. On the other hand, the work (Tung et al. 2016) that adopted this approach to predict election results at municipal level, achieved a partial success even using millions of opinions, i.e., the correct winner was predicted correctly for some cities but not all.

### 3.4 Popularity-based approach

This subsection presents eight works that use social media to analyze which candidate is the most popular assuming that a high popularity is related to victory. A brief description of them is presented as follows.

You et al. (2015) proposed to predict the results of the 2012 US presidential elections and 2014 US House race by using a novel model called competitive vector autoregression (CVAR), which they applied to data from Flickr (a social media to share images). Flickr data were collected during the days in which debates occurred and on the day of

the election, using candidates' names as keywords. CVAR compares the popularity of different candidates, by combining visual and textual data (in order to extract reliable signals and to reduce sampling bias). The features considered include: (i) image metadata features: composed by the title, the description, and tags associated with the image; and (ii) visual features: are based on the true content of the image. Opencv and the library Stasm were used to extract faces and facial features, respectively. The AdaBoost classifier was adopted to classify the faces that were found into the following categories: flattering, unflattering, and neutral; and (iii) they adopt Sentiment140 to infer the sentiment (positive or negative) from the viewers' comments. The authors argued that this method accurately predicted the results of the given election.

Kagan et al. (2015) tried to predict the results of the 2013 Pakistani elections and the 2014 Indian elections based on what they call sentiment diffusion models. The Rensselaer Polytechnic Institute provided the author's tweets containing 31 topics (politicians and political parties) related to the Indian elections. After that, they built a Twitter Indian Election Network (TIEN), which is composed of people who posted these tweets, their followers, and people that they follow. This network was composed of over 23 million tweets from over 16 million Twitter users. To each topic, they computed the sentiment by using an adaptation of the adjective-verb-adverb (AVA) algorithm. A diffusion estimation model is learned from the data in order to detect how the opposition to/support for a candidate is spreading through the network. The diffusion method predicts the sentiment in each user's next tweet related to a given topic using the SVM algorithm. The authors state that the outcomes of the elections were predicted correctly.

Dokoohaki et al. (2015) presented an approach to predict the results of two elections, namely 2014 European parliamentary elections and 2014 Swedish General elections. A total of seven million tweets were collected for 8 months. From these tweets, two million were related to political content, mentioning the elections, debate, a political party or candidate. In order to collect relevant data, tweets were filtered based on the coordinates of the country and based on political-related hashtags. The online popularity of candidates was calculated based on link prediction algorithms, which capture the density of conversations about parties or candidates. The authors argue that a stochastic link mining technique can be applied to detect the dynamism of



interactions, which can reflect on a shift in the popularity of the candidates' accounts. Tweets were filtered based on the coordinates of the country and based on political hashtags and mentions to political accounts. An interaction graph was built whose nodes represent users associated with political, and edges represent that a given user mentioned another one at least on time. This graph reveals how cohesive is the interaction between members of the political parties. The popularity of the candidates and parties is estimated based on the density of the interactions. They concluded that there are similarities between the popularity of the candidates/parties on social media and the vote outcomes.

Wang and Lei (2016) presented an approach that uses sentiment analysis and peer-to-peer ratings among the social network in order to predict the outcomes of the 2014 Taiwanese local election based on data collected from the largest forum of Taiwan during 3 months before the election. More than 26,000 articles and one million ratings were collected. This approach calculates the acceptance/popularity of the candidates according to the rating records of their related articles. The sentences of the articles were separated into a sequence of words. A corpus of online language was built containing the top 1000 frequent words that appear in the sentences, after removing stop words. The volumetric step consists in counting the number of articles that contain the name of a given candidate. A sentiment corpus called NTUSD was used to infer the sentiment of the article. The authors consider that a positive sentiment article related to a candidate with many likes indicates that the candidate has a more general acceptance rather than an acceptance from only the article author. For this reason, the public acceptance score was calculated as the sum of the number of positive ratings of a candidate related article minus the number of negative ratings of a candidate related article. The daily statistics of the three indicators (volume, sentiment, public acceptance) were calculated to capture the temporal dynamics of these indicators. The combination of the volume, sentiment scores, and public acceptance scores indicators was used to develop a regression model to predict the final vote shares of candidates. The proposal was evaluated, taking into account a 3-month observation on a popular online social forum in Taiwan during the 2014 local election campaign. The authors have argued that their results outperformed previous approaches in predicting the final vote winners of the elections and the final vote shares.

Xie et al. (2016) proposed to predict the results of the 2016 Taiwan presidential election using different sources of online information (Twitter, Facebook, Google, and candidates' campaign websites) and offline data from pollsters. Taking into account that samples from online data can be age-biased, their prediction method has considered demographic information to weight the online voting shares and offline results. About 246,893 tweets were collected from

October 1, 2015, to January 16, 2016. The daily average number of 'Likes' per post for each candidate was also calculated. A differential factor of such an approach was to use a signal processing method called Kalman filter to automatically select reliable sources of information, fuse them and predict the daily voting percentages. The search popularity of the candidates was considered in their analysis. By using Google Trends, it is possible to know how a term's search volumes have changed over time considering the average number of times some keywords have been searched for some time period. Additionally, the popularity of the candidates was calculated based on the IP traffic of the candidates' webpages. Their results accurately predicted the outcomes of the 2016 Taiwan presidential election with small error rates less than 3%.

Wang and Gan (2017) presented an approach to predict the 2017 French election results using social media. French tweets were collected from April 24 to May 6 using time tags and keywords such as candidates' names and were classified as positive, negative, or neutral based on the combination of domain knowledge and data analysis. For instance, according to domain knowledge, the words "vote," "win" and "lead" can be interpreted as positive words. On the other hand, "bad," "attack" and "betray" can be interpreted as negative words. A data analysis process selected relevant keywords to sentiment analysis based on the frequency of words in the collected tweets. In this way, the connotation (positive or negative) of the most frequent words in the given political context was manually assigned, and these words were considered features for classifying tweets. The popularity of each candidate was calculated based on the positive or negative rate of tweets related to them. Neutral tweets were considered in the popularity analysis taking into account that they can propagandize the candidates. This approach considers that the election winner will be the most popular candidate. This method achieved a result of about 2% different from the election official result (using data extracted one day before the election), predicting the correct winner.

Wang and Gan (2018) proposed a method to predict the 2017 French presidential election outcomes based on Twitter. Their method applies sentiment analysis and term weighting and selection to predict the popularity of candidates. Tweets mentioning candidates' names and election keywords posted before the elections were gathered and were labeled as positive/negative according to their tags. Keywords are weighted based on both statistics (TF-IDF) and domain knowledge (sentimental meanings of keywords). Therefore, words with the highest scores were considered as important features for election prediction. They use domain knowledge to determine if a keyword is positive/negative and use its TF-IDF to determine the weight of the keyword (which will be a positive score when the domain knowledge says that it is positive, and a negative score, otherwise). A candidate score is computed by summing

**Table 5** Popularity-based approach summary

Paper	Year	Type	Location	Sample	Success
You et al. (2015)	2012	Presidential	USA	N/A	✓
Kagan et al. (2015)	2013, 2014	Presidential	Pakistan, India	23,000,000	✓
Dokoohaki et al. (2015)	2014	Parliamentary, general	Europe, Sweden	2,000,000	N/A
Wang and Lei (2016)	2014	Local	Taiwan	1,000,000	✓
Xie et al. (2016)	2016	Presidential	Taiwan	246893	✓
Wang and Gan (2017)	2017	Presidential	France	N/A	✓
Wang and Gan (2018)	2017	Presidential	France	N/A	N/A
Joseph (2019)	2019	General	India	250,000 <sup>a</sup>	N/A

<sup>a</sup>The number of collected tweets was not explicitly presented in the work in Joseph (2019). However, it was inferred based on the number of tweets collected per day

all the scores of his keywords. The popularity of the candidate is calculated as follows:

$$\text{popularity}(a) = \frac{\text{score}(a)}{(\text{score}(a) + \text{score}(b))},$$

assuming that  $a$  is the candidate whose popularity is being predicted and  $b$  is his adversary. First, based on data on May 2, 2017, the top 100 keywords with highest score for each candidate were used to compute the popularity. After, all keywords were used to predict the popularity. The authors concluded that the strategy that used only the top 100 keywords to predict the popularity was not enough to obtain accurate popularity (when compared to the opinion poll), but the strategy that used all keywords and their score was closer to the polling outcomes. By analyzing tweets collected during the day before the election (May 6, 2017), the percentages achieved using all the keywords were very close to the real outcomes.

Joseph (2019) proposed a methodology to predict the outcomes of the 2019 Indian general elections using the sentiment analysis and Twitter. Every day 5000 tweets for each one of the two most popular parties were collected. Those tweets mention party or party leaders names. The preprocessing removed characters not available in ASCII or Unicode, stop words, punctuations and emojis. The experiments reported in the given paper are only on tweets in English language and having the most number of retweets by the users. TextBlob library is used to classify the sentiment of the tweets and the DT algorithm was adopted. The popularity score for each day was computed, using the following formula:

$$\text{Popularity} = \frac{((0 \times \text{negative tweets}) + \left(\frac{\text{neutral tweets}}{2}\right) + \text{positive tweets})}{\text{total tweets}}$$

This process was carried out for 50 days during the election season. Negative tweets are not given any score. According

to the authors, the predicted outcome is close to the actual outcome and most of the pre-polls.

### 3.4.1 Summary

The general characteristics of the papers presented in this section are summarized in Table 5. The column *paper* indicates the paper that is being analyzed, the column *year* refers to the election year, column *type* indicates the type of the election; the column *location* refers to the election location; the column *sample* refers to the quantity of data collected; finally, the column *success* indicates if the approach achieved success to predict the correct election winner. The term *N/A* is used to indicate that the paper does not present enough information to fill the table field. We can observe that most of the papers using this approach succeeded. However, 3 out of 8 works did not inform if they predicted the correct election winner. The sample size also varied a lot when comparing those works, ranging from thousands to millions of opinions.

### 3.5 Other works

We have chosen not to create approach categories containing only a single paper, and for this reason, we describe in this section the nine works that propose more specific methods for election prediction.

Kalampokis et al. (2017) tried to predict the results of the 2010 UK general election. Their approach assumes that tweets that talk about the UK elections come from

the UK. A total of 84,375 tweets/retweets were collected for 1 month based on a predefined set of terms and hashtags related to the UK elections. A dynamic search

term selection approach was adopted to relate tweets to political parties. Named entities were extracted from the tweets, and existing links among the entities identified (using DBpedia) were explored. These links can be useful to identify that tweets that refer to different entities are referring to the same party. In order to predict election results, the volume of tweets per party is calculated, and sentiment analysis was calculated based on the machine learning classifier (DynamicLMClassifier). The training set was built considering tweets containing a set of negative hashtags (#dontvotetory, #labourout, #libdemfail) or containing positive hashtags (#torywin, #votelabour, #imvotinglibdem). Tweets related to more than one political party were discarded. The preprocessing step included: removal of stop words, user mention entities, and URL entities, phenomenon related hashtags (#ukelection), replace party and candidate names by the same term. A predictive model was developed for each party based on the regression analysis method. The authors state that their result was close to the real outcomes.

Tsakalidis et al. (2015) presented an approach to predict the outcomes of Germany, Netherlands, and Greece elections that took place in 2014. Such an approach treats users' voting intentions as time-variant features. In this way, Twitter-based features are fitted in time-series models and combined with opinion polls, which are adopted as the ground truth. As examples of Twitter-based features and poll-based features, we can cite the number of tweets mentioning a party on a specific day and the percentage of the given party according to a poll of this day, respectively. Twitter data and data from different opinion polls were collected from April 6 up to May 23, filtering by data containing a party name or its abbreviation, mentioning the candidate account. In addition, ambiguous keywords were removed (e.g., the abbreviation of the Dutch party "GL" can be an abbreviation for "good luck"). The number of tweets mentioning a party per day and the number of tweets that have positive/negative sentiment were calculated. A total of 361,713 tweets were collected from 74,776 users in Germany, 452,348 from 74,469 users in the Netherlands, and 263,465 from 19,789 users in Greece. For the sentiment analysis step, a lexicon-based approach was adopted, and to deal with the lack of lexicon for some languages, Google translator was adopted to translate the SentiWordNet, the Opinion Lexicon, and the Subjectivity Lexicon. Several algorithms (linear regression, Gaussian process, and sequential minimal optimization for regression) were applied on each political party separately, using a specific party's features as input (11 Twitter- and one poll-based). The authors stated that they are among the ones that published their predictions for the Greek election before the announcement of the exit polls. Finally, their experiments achieved lower error rates, even when compared to traditional polls.

White (2016) tried to predict the results of the 2015 Canadian elections using Twitter data. About 2,500,821 tweets containing mentions to the party or party leader were collected between May 1, 2014, and June 12, 2014. The authors use the location field text informed in the user profile to determine if the user is from the desired geography. A step to analyze the representativeness of the sample is performed by measuring the demographic information of 5000 Twitter users from Toronto. These 5000 users were manually classified (3 annotations per user). Only classifications that received at least two equal labels were maintained. The final dataset was composed of 3032 Twitter users together with their characteristics (labels received by the manual annotation). They concluded that the demographics of Twitter users almost follow the official census, except that Filipino ethnic and users with age under 14 years or over 65 years are underrepresented. A Vector Autoregression with Exogenous Variables (VARX) model is adopted to forecast the election results. While the input variable is a Twitter feature, the output variable is a time series of the aggregated daily polls for each major party. Six different features were tested: (i) tweet volume (mentions of a candidate); (ii) tweet SoV (unique people mentioning a candidate); (iii) positive volume (positive mentions of a candidate); (iv) positive SoV (unique people positively mentioning a candidate); (v) mean sentiment volume (mean sentiment score for each candidate); and (vi) mean sentiment SoV (person's mean sentiment score for each candidate). Sentiments were calculated by using the word polarity approach, and mentions were normalized according to the total number of tweets in a given day. To predict the 2015 Canadian Federal elections, 34,732,633 tweets collected between January 1, 2015, and October 19, 2015, were considered from 130,816 users. The model correctly predicted the overall Canadian election and the provincial results. On the other hand, traditional polls missed the winner of one province (British Columbia). However, both the polls and the Twitter forecast underestimated the winner in Quebec.

Kassraie et al. (2017) described a method in which tweets about the 2016 US Elections were gathered and the resulting key trends were validated against using Google Trends in order to create a legitimate dataset. The main idea of this approach is to estimate candidates' vote shares instead of an absolute winner of the election, as occurs in the majority of approaches in the literature. About 370,000 tweets containing the candidates' names were collected in a period of 6 months. Tweets containing more than one candidate were discarded. The tweets' text was analyzed, and the most common terms and hashtags were manually grouped into sets of election-relevant terms based on the collective knowledge of election events. For instance, the *terms gun, guns, guncontrol, stopgunviolence*, are grouped into a set represented by the keyword "gun control." The popularity of the keywords

is validated using Google trends. The sentiment analysis of the content of the tweets was performed using the Rsentiment and the SentimentR packages. The election outcomes were predicted by using a Gaussian process regression model that estimates weekly predictions. This method was applied to the US 2016 Elections and predicted Clinton's majority in the popular vote at the beginning of the elections week with a 1% error.

Fano and Slanzi (2017) analyzed tweets to predict the outcomes of a constitutional referendum that took place in Italy in 2016. This referendum was proposed to approve a constitutional reform that would change the distribution of powers between the state and regions. About one million tweets containing hashtags (#referendumcostituzionale, #iovotono, #bastaunno, #iovotosi, #bastaunsi) related to the referendum were collected during 1 month before the election. The LDA algorithm was adopted to perform a topic modeling where frequent topics and keywords were extracted. From the analysis of the topics, the authors find out the most frequent words related to positive or negative sentiments. Most of the topics connected to hashtags in favor of the constitutional reform (#bastaunsi and #iovotosi) are terms connected to positive sentiments (*future, pride and change*); on the other hand, words in the tweets related to the hashtags that are against the reform (#bastaunno and #iovotono) are linked to words with negative sentiments (*fear, complaint, danger and risk*). The percentage of tweets against the constitutional reform was calculated to estimate the outcome of the referendum. The day with the highest number of tweets has a percentage of 62% against the reform, while the actual outcome of the referendum was 59% against the constitutional reform. Then, the election was predicted correctly, with a prediction error of 3%.

Ajito et al. (2017) focused on predicting the ranking of the number of seats acquired in the 48th Japan House of Representatives Member General Election, which took place in 2017, using the mathematical model of hit phenomenon. The authors assume that advertisements (TV, newspapers, magazines, web news, Twitter) and communication (e.g., the recommendation of a third party or a friend on blogs and Twitter) are factors that influence the electors' vote. Three parties were considered, namely POH, CDP, and the LDP. The mathematical model states that the effect of the external advertisement is calculated based on the number of exposures per unit time (day) (exposure time), the type of the media, and a constant representing the strength of the advertisement effect of such a media type on people. The direct communication is based on a conversation between two persons  $i$  and  $j$  and is calculated based on the probability that  $i$  is affected by  $j$ . The indirect influence comes from a conversation from third parties (and not family and friends) on the street, restaurants, information heard in the public transport, rumors found on blogs and Twitter. In this way,

an indirect conversation involves three persons  $i$ ,  $j$ , and  $k$ , where  $i$  hears the conversation between  $j$  and  $k$ . The influence of the conversation between  $j$  and  $k$  is calculated based on the  $i$ 's attention/motivation. The changes in the number of writings for each political party were analyzed before the public announcement of the election until the day before the Election Day. Considering blogs, the number of seats acquired is expected to be the order of LDP > CDP > POH. Although the party CDP has posted a large number of tweets per month, the average impression per tweet is lower than the LDP. The same occurs for the POH. The authors have argued that the positive influence of election activities by Twitter is low, and the positive influence of the blog is high, since the ranking obtained with the blog analysis is equal to the real ranking of the number of seats acquired, suggesting that a mathematical model of hit phenomenon can be used for national elections.

Huang (2017) presented a study to predict the outcomes of the 2014 Taiwan mayoral election by analyzing data related to the two main candidates. Their research is divided into the following subtasks: (i) identify the election related attribute words manually; (ii) identify the opinions related to the attribute words; (iii) analyze the polarity of the opinions using the National Taiwan University Sentiment Dictionary (NTUSD)—a lexical released by Taiwan university; (iv) determine the strength of adverbs using the HowNet, which rates these words from 1 to 5; (v) determine the negative words using a predefined list of 25 negative words—if a negative word is found in the sentence, its polarity is changed to be the opposite, i.e., negative sentences become positive and vice versa; (vi) determine the score of the opinions based on their polarity and strength. A total of 4419 opinions were collected from January 1 to November 28 in 2014 as follows: 2009 from e-news; 167 from magazines; and 2243 from Facebook. Sentences with neutral opinion are discarded. Candidates percentages were predicted by summing the scores for each topic and this approach was able to predict the election winner.

Awais et al. (2019) proposes to predict the Pakistan's 2018 General Election using Twitter. Instead of predicting overall vote share of major political parties, they propose to predict the winning probability of the candidates for each constituency. They employed a Bayesian optimization model that combines three types of data: results of the four past elections—to capture party influence; public poll data of the last 2 years—to capture popularity levels of major political parties; and tweets of 3 weeks before the election related to the four major political parties—to capture candidates popularity. Each data source provides a probability vector for each constituency. A total of 640,000 tweets were collected using keywords related to party and candidates names during 18 days before the Election Day, storing its text, number of retweets, and the number of favorites it got.

**Table 6** Other works summary

Paper	Year	Type	Country	Sample	Success
Kalampokis et al. (2017)	2010	General	United Kingdom	84,375	N/A
Tsakalidis et al. (2015)	2014	Parliamentary	Germany, Netherlands, Greece	361,713, 452,348, 263,465	N/A
White (2016)	2015	Federal	Canada	34,732,633	Partial
Kassraie et al. (2017)	2016	Presidential	USA	370,000	No
Fano and Slanzi (2017)	2016	Referendum	Italy	N/A	✓
Ajito et al. (2017)	2017	General	Japan	N/A	Partial
Huang (2017)	2014	Mayoral	Taiwan	4419	✓
Awais et al. (2019)	2018	General	Pakistan	640,000	Partial
Brito and Adeodato (2020)	2018, 2016	Presidential, presidential	Brazil, USA	252,000,000, N/A	✓

Sentiment analysis was performed to classify tweets, where a score equals to: 1 shows extreme positive emotions; -1 shows extreme negative emotions; and 0 shows a neutral sentiment. The party popularity is computed by the following formula:  $p_s = p \times (0.02f_c + 0.01r_c)$ , where  $p_s$ : popularity score,  $p$ : sentiment polarity,  $f_c$ ,  $r_c$ : favorite and retweet counts.<sup>5</sup> This work does not inform the sentiment analysis software that was adopted. This prediction strategy was also able to predict the winner correctly for 150 out 270 seats.

Brito and Adeodato (2020) presented analyses about the 2018 Brazilian presidential election and the 2016 US presidential election. Their analysis focuses on the repercussion of the posts of official candidates' profiles considering Facebook, Twitter, and Instagram. They combine the sum of likes, shares and comments of social media posts with traditional polls to train ML models to predict vote share to each candidate individually. Therefore, the result of a poll at a specific date is a function of the repercussion observed in the candidate social media during an aggregate window of days prior the given poll date. Two algorithms were adopted: a multilayer perceptron (MLP) and, for baseline, traditional linear regression. In regard to the Brazilian elections, social media data from the 5 most popular candidates were collected and polling data that were published by Ibope and Datafolha from January 1, 2018, until the day before the election, considering a total of 21 polls. These Brazilian candidates made a total of 18,976 posts during the analyzed period, which resulted in 252 million interactions. This approach was able to predict the election winner correctly using both LR and MLP algorithms. MLP achieved the best vote share in relation to real outcomes. In the US election experiment, the aim of the authors is to predict the final popular vote share of each one of the two candidates (and not to predict the elected candidate as the US president election process is indirect). For the US election, polling data were

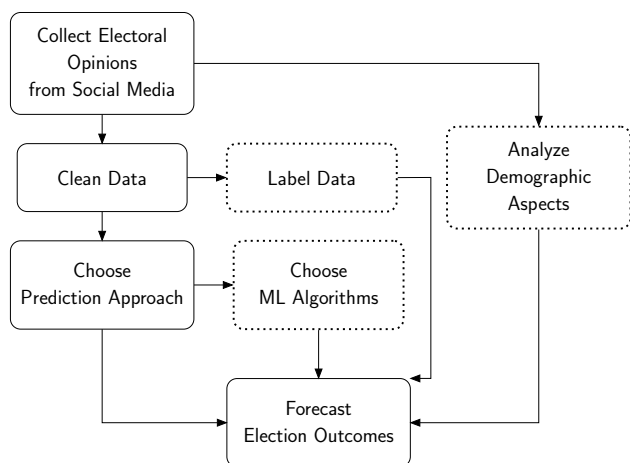
collected starting from 1 year before elections, November 08, 2015, until November 07, 2016. The vote share results obtained with this experiment were inline with the real outcomes and were better than the results obtained using only traditional polls.

### 3.5.1 Summary

The general characteristics of the papers presented in this section are summarized in Table 6. The column *paper* indicates the paper that is being analyzed, the column *year* refers to the election year, column *type* indicates the type of the election; the column *country* refers to the election country; the column *sample* refers to the quantity of data collected; finally, the column *success* indicates if the approach achieved success to predict the correct election winner. The term *N/A* is used to indicate that the paper does not present enough information to fill the table field. The works in White (2016) and Awais et al. (2019) are classified as partial success since their methods were able to predict the winner correctly only for some cities/seats. Similarly, the method presented in Ajito et al. (2017) is classified as partial success as the winner ranking built based on blog content was inline with election results but the winner ranking built based on Twitter did not reflect election results.

In this section, we presented a taxonomy for grouping works according to the approaches we identified they follow. This is an important contribution for those who want to analyze electoral scenarios using social media. Another point is that there is a lack of frameworks to help data scientists to guide data analysis in this specific scenario. These frameworks are interesting to help guaranteeing key aspects of trustworthy AI, which includes replicability, reliability, auditability, transparency and accountability (Janssen et al. 2020). In this way, next section presents the main tasks used in the data science processes followed by the analyzed works.

<sup>5</sup> The constants with  $f_c$  and  $r_c$  were chosen based on trial and error.



**Fig. 2** Sketch of a data and opinion mining process for election outcome prediction using social media data

#### 4 A general opinion mining process for social media analysis to elections outcomes predictions

In this section, we analyze the works described in the previous section according to the steps they follow in their data and opinion mining processes. We identified the following main tasks that compose the general process: data collection details (social media, amount of collected data, collection period, keywords used to collect data), preprocessing steps, strategies for labeling data—in the cases that it is needed (e.g., supervised methods), machine learning algorithms used, demographic information that was adopted, and the approach that they follow for forecasting elections outcomes. Figure 2 presents a sketch of this general process. In this figure, rectangles with rounded corners represent tasks, dotted borders rectangles represent optional tasks and the arrows relate a predecessor task to a successor task. This sketch presents the tasks usually adopted by the papers, briefly described in what follows. Each one of these main tasks are deeply analyzed in the following subsections.

- **Collect Electoral Opinions from Social Media**—this task is the first one to start the electoral analysis process and refers to all decisions and actions related to *data collection*, namely choose which social media the data will be collected from; choose what is the collection period and/or choose what keywords will be used to collect data; and collect data;
- **Clean Data**—this task is the second one in the data analysis process and refers to the data *preprocessing* steps that are conducted to clean data;
- **Choose Prediction Approach**—this task refers to the *approach* that the papers will follow for forecasting election outcomes (e.g., Counting-Based Approach, Political

Alignment Approach, Event Detection Approach, Popularity-Based Approach, or other approach);

- **Label Data**—choose and apply strategies for labeling data after cleaning data. This task is optional and is only needed by supervised approaches, which are the ones that require labeled datasets;
- **Analyze Demographic Aspects**—refers to the *demographic information* that will be considered in the analysis, such as user age, gender or location. This task is optional and can be used to weight the contribution of opinions in their analysis, filter opinions by location, or remove non-representative users, for example;
- **Choose Machine Learning Algorithm**—this task is optional and refers to the action of choosing what *machine learning algorithms* will be used in classification or regression predictions. For instance, the Counting-Based Approach may require sentiment analysis classification that can be performed using machine learning techniques. Similarly, the Political Alignment Approach requires the classification of users into different political categories;
- **Forecast Election Outcomes**—this is the final task of the process and it is different according to the prediction approach that was adopted. For example, if the Popularity-Based Approach was adopted, then, probably a formula to predict popularity will be proposed to predict the election winner. On the other hand, if the Event Detection Approach is selected, then, a rule-based strategy will be used to compute the election winner based on the detected events (see Sect. 3 for more details related to the specific strategies used by each paper to predict election outcomes). This task can also use the labeled data, execute machine learning algorithms and use content related to demographic aspects to obtain better results.

##### 4.1 Data collection

This section refers to information about how data were collected from social media, such as data sources, quantity of data collected, keywords used for gathering data and collection period. Election data to be analyzed are collected from a given data source and usually are collected based on a given time period or based on keywords/search terms. We observed that the collection period was variable ranging from less than a month to more than 6 months. In relation to the search keywords, we identified that most of the papers use keyword of the following categories: (i) *candidate related*: terms or hashtags including candidate name or last name; (ii) *party names*: term or hashtags that refer to party names; and (iii) *election keywords*: terms or hashtags containing campaign slogans, for example. A summary about the opinion sources is presented as a Venn diagram in Fig. 3. Concerning data sources, we notice that

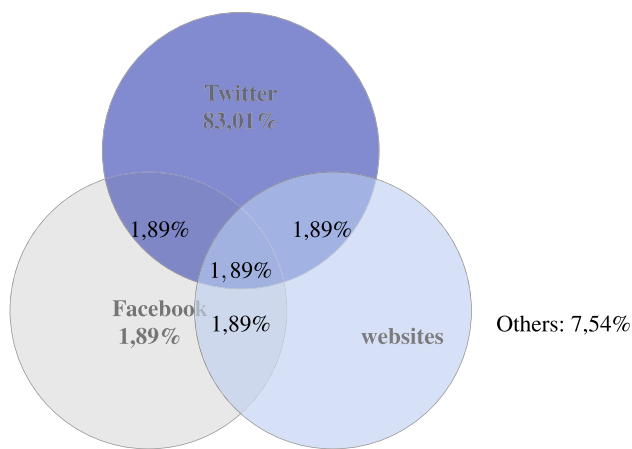


Fig. 3 Opinion sources

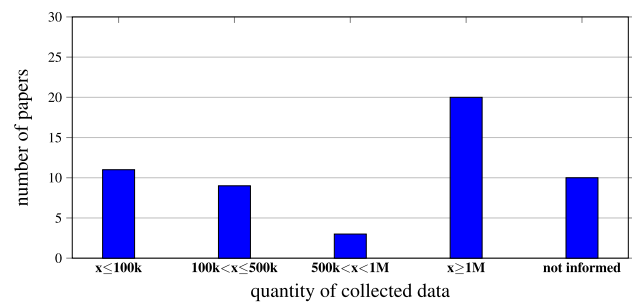


Fig. 4 Number of papers by quantity of collected data

Table 7 refers to the amount of data collected. The column *paper* is the reference to the paper in which the approach was detailed; the column *number of posts (x)* refers to the num-

Table 7 Quantity of data instances collected

Number of posts (x)	Paper
$x \leq 100,000$	Ramzan et al. (2017), Sharma and Moh (2016), Maldonado and Sierra (2015), Jose and Chooralil (2016), Kalam-pokis et al. (2017), Hinch (2017), Singh et al. (2020), Kristiyanti et al. (2019), Bilal et al. (2018), Singh et al. (2017) and Huang (2017)
$100,000 < x \leq 500,000$	Kassraie et al. (2017), Xie et al. (2016), Tsakalidis et al. (2015), Praciano et al. (2018), Joseph (2019) <sup>a</sup> , Lopardo and Brambilla (2018), Sanders et al. (2016), Bansal and Srivastava (2018), Bansal and Srivastava (2019)
$500,000 < x < 1,000,000$	Castro and Vaca (2017), Unankard et al. (2014), Awais et al. (2019)
$x \geq 1,000,000$	Srivastava et al. (2015), Burnap et al. (2016), Shaban et al. (2017), Khatua et al. (2015), Dokoohaki et al. (2015), Almeida et al. (2015), Rosseti et al. (2017), Ibrahim et al. (2015), Heredia et al. (2017), Fano and Slanzi (2017), Dwi Prasetyo and Hauff (2015), Campanale and Caldarella (2018), Wang and Lei (2016), Vepsäläinen et al. (2017), Kagan et al. (2015), Tung et al. (2016), Heredia et al. (2018), White (2016), Bastos and Mercea (2018) and Brito and Adeodato (2020)
Not informed	Ajito et al. (2017), Wang and Gan (2017), You et al. (2015), Wicaksono et al. (2016), Wang and Gan (2018), Naiknaware and Kawathekar (2018), Bachhuber et al. (2016), Sanders and van den Bosch (2020), Budiharto and Meiliana (2018), Hwang (2019)

<sup>a</sup>The quantity of tweets collected was not explicitly presented in this paper (Joseph 2019). However it was inferred based on the number of tweets collected per day

Twitter stands out among social networks for gathering political opinions in order to forecast election, i.e., considering all the 53 papers (100%) analyzed in this research, 44 papers (83.01%) use only Twitter as source of social media election opinions, 1 (1.89%) uses Facebook as the only source of opinions, 1 (1.89%) combines data from Facebook, Twitter and websites (candidates webpages and Google), 1 (1.89%) combines data from Facebook, Twitter, Instagram, traditional polls and past elections, 1 (1.89%) combines data from Facebook and websites (e-news and magazines), and 1 (1.89%) uses data from Twitter and websites (blogs). Finally, 4 papers (7.55%) use exclusively other sources for mining opinions that are not adopted by other papers such as the Flickr, Reddit, the BSS, and a Taiwan forum. We have chosen not to illustrate the latter case in the diagram because these sources are not mentioned by more than one paper.

ber of data instances collected to be analyzed (e.g., tweets, Facebook likes or comments). The ranges of data collection were organized as follows:

- $x \leq 100,000$ : papers that collected up to 100,000 data instances;
- $100,000 < x \leq 500,000$ : papers that collected between 100,000 and 500,000 data instances;
- $500,000 < x < 1,000,000$ : papers that collected more than 500,000 and less than 1 million data instances;
- $x \geq 1,000,000$ : papers that collected more than 1 million data instances.

Papers that do not explicitly inform how much instances were collected are grouped into the *not informed* field (see Table 7). Figure 4 illustrates this information using a bar

**Table 8** Number of successful approaches according to the amount of collected data

	Success	Partial	No	N/A
$x \leq 100k$	6	1	2	2
$100k < x \leq 500k$	3	3	1	2
$500k < x < 1 \text{ M}$	2	1	0	0
$x \geq 1 \text{ M}$	9	5	2	4

chart where we can see that most works collect more than one million data. Table 8 exhibits a summary about the number of successful approaches according to the amount of data, where each row represents the number of papers of a given amount of data range that are associated with each one of the following possibilities: *success*—the paper predicted correctly the election winner in all their experiments; *partial*—the paper achieved success in predicting the election winner in at least one of their experiments but not in all experiments; *no*—the paper failed to predict the election winner; *N/A*—the paper does not present enough information about the success of their approach.

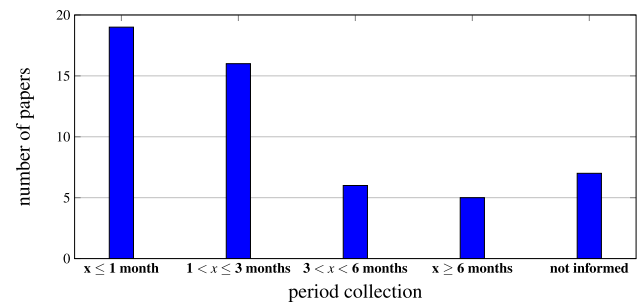
Although it may seem that collecting more data leads to better results, we cannot draw this conclusion, as the amount of works that collected more than 1 million posts is also much greater than the amount of works in the other data collection ranges.

Table 9 presents information about the period in which the data were collected. The column *paper* refers to the paper in which the approach is described; the column *period collection* refers to the period (number of months ( $x$ )) that each approach considered to collect data. The ranges of the period collection were organized as follows:

- **$x \leq 1$  month:** papers that collected data in a period up to 1 month;
- **$1 < x \leq 3$  months:** papers that collected data in a period between 1 and 3 months;
- **$3 < x < 6$  months:** papers whose period of data collection was between 3 and 6 months.
- **$x \geq 6$  months:** papers that collected data in a period bigger than 6 months;

Papers that do not explicitly inform what was the period of data collection are grouped into the *not informed* (see Table 9). Figure 5 illustrates an overview about the collection period using a bar chart where we can see that most works adopt a period between 1 and 3 months.

We have observed that the collection period does not necessarily implies on a higher amount of data (as is the case of the papers Vepsäläinen et al. 2017; Srivastava et al. 2015;

**Fig. 5** Number of papers by period collection

Rosseti et al. 2017; Fano and Slanzi 2017 in Tables 7 and 9, for example). Therefore, the amount of data also depends of the hashtags used for data collection.

Table 10 exhibits a summary about the number of successful approaches according to the data collection period, where each row represents the number of papers of a given data collection period that are associated with each one of the following possibilities: *success*—the paper predicted correctly the election winner in all their experiments; *partial*—the paper achieved success in predicting the election winner in at least one of their experiments but not in all experiments; *no*—the paper failed to predict the election winner; *N/A*—the paper does not present enough information about the success of their approach. While it might seem that collecting data for a shorter time results in better predictions, we cannot draw this conclusion as most works have adopted a short data collection time.

The keywords/terms used to collect data are summarized in Table 11. We observed that keywords related to candidates such as those that use parts of the candidate's first or last name and keywords related to election terms such as the ones that contain campaign slogans or combinations mentioning the name of the elections and the election year are the most popular types of keywords.

## 4.2 Data preprocessing

As pointed out by Liu (2020), social media data are very noisy since they include different kinds of spelling, punctuation and grammatical errors. For this reason, before data analysis, it is important to conduct a preprocessing phase to clean data and remove noise. Table 12 exhibits a summary of the most used preprocessing techniques that were adopted after the data collection phase. In Table 12, the term *word extension* refers to words with duplicated letters such as *Loooove* instead of *Love*.

We have observed that the removal of user mentions, URLs, punctuation, and stop words are the most popular preprocessing steps. A minimal number of works discard duplicated content or try to detect and discard content posted



**Table 9** Period collection

Period collection (x)	Paper
$x \leq 1$ month	Ramzan et al. (2017), Wicaksono et al. (2016), Sharma and Moh (2016), Jose and Chooralil (2016), Srivastava et al. (2015), Vepsäläinen et al. (2017), Rosseti et al. (2017), Kalampokis et al. (2017), Fano and Slanzi (2017), Praciano et al. (2018), Campanale and Caldarola (2018), Singh et al. (2020), Singh et al. (2017), Hwang (2019), Bansal and Srivastava (2018), Sanders et al. (2016), Bansal and Srivastava (2019), Awais et al. (2019), Sanders and van den Bosch (2020)
$1 < x \leq 3$ months	Castro and Vaca (2017), Heredia et al. (2017), Almeida et al. (2015), Ibrahim et al. (2015), Khatua et al. (2015), Tsakalidis et al. (2015), Dwi Prasetyo and Hauff (2015), Joseph (2019), Maldonado and Sierra (2015), Heredia et al. (2018), Unankard et al. (2014), White (2016), Bastos and Mercea (2018), Shaban et al. (2017), Wang and Lei (2016), Bilal et al. (2018)
$3 < x < 6$ months	Xie et al. (2016), Burnap et al. (2016), Tsakalidis et al. (2015), Tung et al. (2016), Kassraie et al. (2017) and Budiharto and Meiliana (2018)
$x \geq 6$ months	Dokoohaki et al. (2015), Lopardo and Brambilla (2018), Hinch (2017), Brito and Adeodato (2020) and Huang (2017)
Not informed	Ajito et al. (2017), You et al. (2015), Kagan et al. (2015), Wang and Gan (2018), Kristiyanti et al. (2019), Bachhuber et al. (2016) and Naiknaware and Kawathekar (2018)

**Table 10** Number of successful approaches according to the data collection period

	Success	Partial	No	N/A
$x \leq 1$ month	9	5	1	4
$1 < x \leq 3$ months	8	4	1	3
$3 < x < 6$ months	3	1	1	1
$x \geq 6$ months	2	1	1	1

### 4.3 Data labeling

The methods that were adopted by the papers for labeling data are presented in Table 13. Figure 6 summarizes the information about the labeling methods using a Venn diagram. From the 38 papers that *use methods for predicting sentence sentiment*, three of them (Ramzan et al. 2017; Singh et al. 2020; Awais et al. 2019) do not inform the method that was adopted to assign polarities to the sentences and for this

**Table 11** Types of keywords used by the surveyed papers

Keywords	Paper
Candidate related	Heredia et al. (2017), Jose and Chooralil (2016), Dwi Prasetyo and Hauff (2015), Burnap et al. (2016), Wicaksono et al. (2016), Khatua et al. (2015), Heredia et al. (2018), Maldonado and Sierra (2015), Ibrahim et al. (2015), Rosseti et al. (2017), Shaban et al. (2017), You et al. (2015), Wang and Gan (2017), Dokoohaki et al. (2015), Kassraie et al. (2017), White (2016), Tsakalidis et al. 2015, Praciano et al. (2018), Joseph (2019), Wang and Gan (2018), Singh et al. (2020), Kristiyanti et al. (2019), Bilal et al. (2018), Hinch (2017), Singh et al. (2017), Bansal and Srivastava (2018), Bansal and Srivastava (2019), and Awais et al. (2019)
Party names	Sharma and Moh (2016), Burnap et al. (2016), Wicaksono et al. (2016), Khatua et al. (2015), Maldonado and Sierra (2015), Dokoohaki et al. (2015), White (2016), Tsakalidis et al. (2015), Joseph (2019), Singh et al. (2020), Bilal et al. (2018), Singh et al. (2017), Bansal and Srivastava (2018), Bansal and Srivastava (2019), Awais et al. (2019) and Sanders and van den Bosch (2020)
Election keywords	Wicaksono et al. (2016), Khatua et al. (2015), Heredia et al. (2018), Tung et al. (2016), Unankard et al. (2014), Shaban et al. (2017), Dokoohaki et al. (2015), Kagan et al. (2015), Fano and Slanzi (2017), Kalampokis et al. (2017), Campanale and Caldarola (2018), Wang and Gan (2018), Kristiyanti et al. (2019), Naiknaware and Kawathekar (2018), Hinch (2017), Bansal and Srivastava (2018), Bansal and Srivastava (2019), Budiharto and Meiliana (2018), Huang (2017)
Not informed/ do not use key- words	Ramzan et al. (2017), Srivastava et al. (2015), Xie et al. (2016), Wang and Lei (2016), Ajito et al. (2017), Castro and Vaca (2017), Almeida et al. (2015), Bastos and Mercea (2018), Vepsäläinen et al. (2017), Sanders et al. (2016), Bachhuber et al. (2016), Sanders and van den Bosch (2020), Brito and Adeodato (2020) and Hwang (2019)

by bot accounts (spam). Most of the works did not inform the steps conducted for data preprocessing. A few works translate opinions during preprocessing step. Techniques to filter bots and non-personal accounts are little explored.

reason were not considered in this analysis. From the remaining 35 papers (100%), the total of 17 papers (48.57%) rely only on lexicon dictionaries to determine the sentiment of a sentence. Methods that rely only on emoticon or hashtags that denote positivity/negativity are also trendy (22.86%) (8 papers). Only three papers (8.57%) are based on manually labeling a subset of the documents (semi-supervised

**Table 12** Preprocessing steps

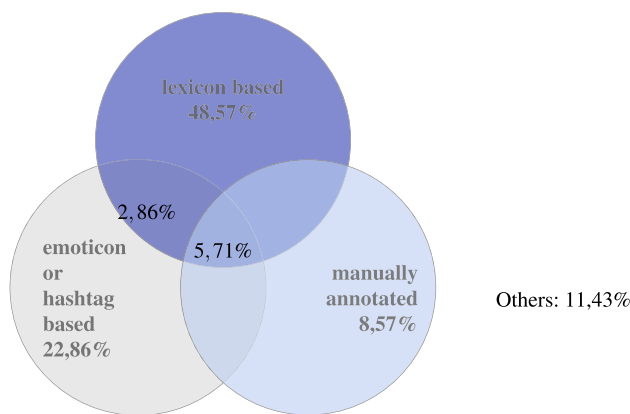
Preprocessing step	Papers
Lower case conversion	Ramzan et al. (2017), Praciano et al. (2018), Singh et al. (2020), Naiknaware and Kawathekar (2018), Bansal and Srivastava (2018) and Bansal and Srivastava (2019)
Hashtags removal	Ramzan et al. (2017), Sharma and Moh (2016), Praciano et al. (2018), Campanale and Caldarola (2018), Naiknaware and Kawathekar (2018) and Singh et al. (2017)
Stop words removal	Ramzan et al. (2017), Sharma and Moh (2016), Wicaksono et al. (2016), Ibrahim et al. (2015), Castro and Vaca (2017), Tung et al. (2016), Unankard et al. (2014), Praciano et al. (2018), Campanale and Caldarola (2018), Joseph (2019), Singh et al. (2020), Naiknaware and Kawathekar (2018), Bansal and Srivastava (2018), Bansal and Srivastava (2019) and Budiharto and Meiliana (2018)
URL removal	Ramzan et al. (2017), Sharma and Moh (2016), Wicaksono et al. (2016), Khatua et al. (2015), Ibrahim et al. (2015), Unankard et al. (2014), Praciano et al. (2018), Singh et al. (2020), Naiknaware and Kawathekar (2018), Bilal et al. (2018), Singh et al. (2017) and Budiharto and Meiliana (2018)
Emoticons removal	Heredia et al. (2017), Sharma and Moh (2016), Castro and Vaca (2017), Bilal et al. (2018)
duplicated removal	Heredia et al. (2017), Khatua et al. (2015), Bansal and Srivastava (2019) and Budiharto and Meiliana (2018)
Retweets removal	Heredia et al. (2017) and Sharma and Moh (2016)
Special characters removal	Sharma and Moh (2016), Wicaksono et al. (2016), Joseph (2019), Budiharto and Meiliana (2018), Kristiyanti et al. (2019) and Bansal and Srivastava (2018)
Punctuation removal	Sharma and Moh (2016), Ibrahim et al. (2015), Castro and Vaca (2017), Tung et al. (2016), Praciano et al. (2018), Joseph (2019), Singh et al. (2020), Kristiyanti et al. (2019), Naiknaware and Kawathekar (2018), Bilal et al. (2018), Bansal and Srivastava (2018) and Bansal and Srivastava (2019)
Username mentions removal	Ramzan et al. (2017), Sharma and Moh (2016), Wicaksono et al. (2016), Ibrahim et al. (2015), Unankard et al. (2014), Praciano et al. (2018) and Naiknaware and Kawathekar (2018)
Negation handling	Sharma and Moh (2016), Jose and Chooralil (2016) and Unankard et al. (2014)
Ambiguous keywords removal	Tsakalidis et al. (2015)
Bots and non-personal accounts removal	Almeida et al. (2015), Ibrahim et al. (2015), Castro and Vaca (2017) and Dwi Prasetyo and Hauff (2015)
Stemming	Castro and Vaca (2017), Unankard et al. (2014) and Bansal and Srivastava (2018)
Retweet keyword (RT)	Unankard et al. (2014), Ibrahim et al. (2015), Naiknaware and Kawathekar (2018) and Bilal et al. (2018)
Slang and word extension replacement	Unankard et al. (2014) and Naiknaware and Kawathekar (2018)
Translation	Singh et al. (2020) and Praciano et al. (2018)
Not informed	Jose and Chooralil (2016), Dwi Prasetyo and Hauff (2015), Burnap et al. (2016), Srivastava et al. (2015), Maldonado and Sierra (2015), Heredia et al. (2018), Vepsäläinen et al. (2017), Rosseti et al. (2017), Bastos and Mercea (2018), Shaban et al. (2017), You et al. (2015), Wang and Gan (2017), Xie et al. (2016), Dokooohaki et al. (2015), Kagan et al. (2015), Wang and Lei (2016), Kassraie et al. (2017), Ajito et al. (2017), Fano and Slanzi (2017), White (2016), Wang and Gan (2018), Hinch (2017), Sanders et al. (2016), Bachhuber et al. (2016), Sanders and van den Bosch (2020), Brito and Adeodato (2020), Awais et al. (2019), Huang (2017) and Hwang (2019)

approach) and 5.71% (i.e., 2 papers) of the works combine all of these three methods (lexicon, emoticon/hashtag and manually annotated). One paper (2.86%) uses emoticons and lexicons to determine sentence sentiment. A total of 11,43% of the works (4 papers) use other methods that are adopted by only one approach, and, for this reason, we choose not to illustrate them in the diagram (which exhibits information about the major three methods). As an example of this case is the SAS software cited by Maldonado and Sierra (2015) and the Aylien API adopted by Hwang (2019), whose underlying methods to assign polarities are not explained.

The exploratory study about the sentiment analysis process of the 2018 Brazilian presidential elections (Santos et al. 2021) suggests that generic sentiment labeling methods may not be enough to capture the real sentiment of electoral tweets and points out that the use of automatic labeling strategies can be a threat to obtain reliable electoral analyses based on social media. This work compares labels obtained with Microsoft Azure Sentiment Analysis API with labels obtained from manual labeling based on crowdsourcing with the majority voting strategy. Such study showed that the overall sentiment (positive, negative, neutral) of the sample of tweets obtained with the

**Table 13** Data labeling methods

Method	Papers
Emoticons/hashtags based	Heredia et al. (2017), Dwi Prasetyo and Hauff (2015), Srivastava et al. (2015), Wicaksono et al. (2016), Almeida et al. (2015), Heredia et al. (2018), You et al. (2015), Kalampokis et al. (2017), Campanale and Caldarola (2018), Wang and Gan (2018), Kristiyanti et al. (2019), Bansal and Srivastava (2019)
Manually labeled	Sharma and Moh (2016), Srivastava et al. (2015), Almeida et al. (2015), Naiknaware and Kawathekar (2018), Sanders and van den Bosch (2020)
Lexicon based	Tsakalidis et al. (2015), Jose and Chooralil (2016), Burnap et al. (2016), Srivastava et al. (2015), Khatua et al. (2015), Almeida et al. (2015), Ibrahim et al. (2015), Wang and Gan (2017), Wang and Lei (2016), Fano and Slanzi (2017), White (2016), Unankard et al. (2014), Shaban et al. (2017), Praciano et al. (2018), Joseph (2019), Singh et al. (2017), Bansal and Srivastava (2018), Bansal and Srivastava (2019), Budiharto and Meiliana (2018) and Huang (2017)
Others	Maldonado and Sierra (2015), Kagan et al. (2015), Bilal et al. (2018) and Hwang (2019)
Not informed	Ramzan et al. (2017), Singh et al. (2020) and Awais et al. (2019)



**Fig. 6** Labeling methods

automatic labeling strategy was different from the overall sentiment calculated using the manual labeling strategy.

#### 4.4 Demographic information

Table 14 summarizes aspects related to the user profile and location. The column *user* is checked when the approach considers information related to the user profile in its analysis (e.g., sex, age, education, income, etc.). The column *location* is checked when the approach tries to detect the location of the post or of the user, as for example: the work in Unankard et al. (2014)—that uses POS-tag information to identify the location; the approach in White (2016)—that uses the Twitter geolocation tag; the study described in Heredia et al. (2018)—that searches for the location information in the user profile; or even the approach in Sharma and Moh (2016)—that assumes that tweets written in Hindi language belong to Hindi users, and so on. The symbol “-” indicates

that such approach does not consider user characteristics/ location information in its analysis. Figure 7 presents this information in a visual way. We have observed that 62.26% of the works (33 papers) do not consider location and user characteristics info, 7.55% of them (4 papers) consider only user characteristics, and 22.64% (12 papers) consider only location info.<sup>6</sup> A total of 7.55% (4 papers) of the works consider both location and user characteristics. The work in Sanders et al. (2016) and Almeida et al. (2015) infer automatically age and gender of the users based on the history of posts of the user. In addition, Almeida et al. (2015) also uses a name dictionary to infer user gender and a classifier to predict user social class. The work in Bansal and Srivastava (2018) and Bansal and Srivastava (2019) only applied geotagging during data collection when using keywords that are not exclusive to the given election.

#### 4.5 Machine learning methods

Table 15 is related to the machine learning methods that were applied in each election forecast approach. The column *paper* refers to the paper in which the approach was described and the column *machine learning method* refers to the machine learning algorithm used. The column *success rate* indicates the percentage of success of each algorithm considering all works that had success when using this method and the total of works that use it, disregarding works that do not explicitly inform that they predicted the winner of the elections correctly (N/A). In the case where the algorithm is only associated with works that did not achieve success when adopting such algorithm, the success rate is 0%. Also, the symbol “-” is used to indicate cases where the algorithm is only associated with works whose success

<sup>6</sup> From this set of papers, Bansal and Srivastava (2018) and Bansal and Srivastava (2019) use location when keywords are not exclusive to elections.

**Table 14** Demographic information

Paper	User	Location
Ramzan et al. (2017), Heredia et al. (2017), Jose and Chooralil (2016), Burnap et al. (2016), Srivastava et al. (2015), Maldonado and Sierra (2015), Tung et al. (2016), Shaban et al. (2017), You et al. (2015), Wang and Gan (2017), Kalampokis et al. (2017), Fano and Slanzi (2017), Kagan et al. (2015), Wang and Lei (2016), Kassraie et al. (2017), Ajito et al. (2017) and Tsakalidis et al. (2015)	–	–
Sharma and Moh (2016), Wicaksono et al. (2016), Khatua et al. (2015), Heredia et al. (2018), Rosseti et al. (2017), Bastos and Mercea (2018), Unankard et al. (2014), Dokoohaki et al. (2015), Lopardo and Brambilla (2018), Bansal and Srivastava (2018), Hinch (2017) and Bansal and Srivastava (2019)	–	✓
Ibrahim et al. (2015), Vepsäläinen et al. (2017), Xie et al. (2016) and Sanders et al. (2016)	✓	
Dwi Prasetyo and Hauff (2015), Almeida et al. (2015), Castro and Vaca (2017) and White (2016)	✓	✓

is not informed (N/A). Papers that did not adopt machine learning methods are not mentioned in this table. The work in Sanders et al. (2016) adopted a software to automatically users demographic characteristics (age and gender) but it is not clear if such software uses machine learning methods.

#### 4.5.1 Classification tasks

Several classification tasks were adopted in the surveyed approaches such as the detection of buzzer/spammer accounts, demographic info classification (gender, social class, age), political alignment classification, and sentiment analysis. We observed that SVM, Naive Bayes, and Decision Trees are the most common machine learning methods used to address this task. The works in Ramzan et al. (2017) and Maldonado and Sierra (2015) do not appear in Table 15 because it is not clear in these papers if the sentiment analysis is performed using machine learning.

Regarding deep learning, we observed that works using this kind of models are becoming more popular in recent years, i.e., since 2017. The deep learning methods adopted in the surveyed papers are as follows: convolutional neural networks (CNN)—adopted by papers published in 2017, namely Heredia et al. (2017) and Shaban et al. (2017), and Heredia et al. (2018), which adopted an AlexNet model, specifically, and was published in 2018; recurrent neural networks (RNN)—adopted by papers published in 2018, namely Bilal et al. (2018) and Lopardo and Brambilla

(2018), which adopted a RNN-LSTM model, specifically. The paper (Bastos and Mercea 2018), which was published in 2018, used deep learning but did not inform the specific algorithm adopted. If we had created a specific category for deep learning, the success rate would be 40%.

#### 4.5.2 Topic modeling

We observed that the LDA was the most popular algorithm to address topic modeling, as it was adopted in Bansal and Srivastava (2018), Tung et al. (2016), Castro and Vaca (2017) and Fano and Slanzi (2017). The BTM model was also adopted in Bansal and Srivastava (2018) to address this task.

#### 4.6 Approaches for predicting election outcomes

Figure 8 presents a pie chart that illustrates the percentage of papers that belong to each approach presented in Sect. 3. Although we have identified different approaches for handling the election prediction task such as analyzing candidate popularity, detecting events that are important for the course of campaigns, and analyzing user political alignment, we can clearly notice that the approach based on counting instances (considering sentiment or not) are still the most adopted strategy for forecasting election outcomes in the literature.

#### 4.7 General remarks

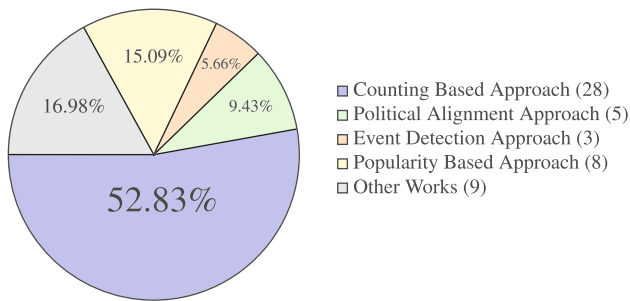
Table 16 presents the general characteristics of the papers. It is organized as follows: the column *ap.* refers to the approach used by the paper to forecast election outcomes. We use numbers to distinguish the different approaches, as follows:

- 1—Counting-Based Approach, described in Sect. 3.1;
- 2—Political Alignment Approach, described in Sect. 3.2;
- 3—Event Detection Approach, described in Sect. 3.3;
- 4—Popularity-Based Approach, described in Sect. 3.4;
- 5—Other Works, described in Sect. 3.5.

**Fig. 7** Demographic information

**Table 15** Machine learning methods

Paper	Machine learning method	Success rate
Srivastava et al. (2015), Sharma and Moh (2016), Castro and Vaca (2017), Almeida et al. (2015), Kagan et al. (2015), Praciano et al. (2018) and Kristiyanti et al. (2019)	Support vector machine	57.4%
Sharma and Moh (2016), Dwi Prasetyo and Hauff (2015), Praciano et al. (2018), Singh et al. (2020)	Naive Bayes	75%
Wicaksono et al. (2016)	Binary multinomial Naive Bayes	100%
Khatua et al. (2015)	OLS Regression	100%
Jose and Chooralil (2016)	Hidden Markov model	–
You et al. (2015)	AdaBoost, CVAR	100%
Almeida et al. (2015)	Random forest	0%
Almeida et al. (2015) and Brito and Adeodato (2020)	Multilayer perceptron	50%
Almeida et al. (2015), Campanale and Caldarola (2018)	Multinomial Naive Bayes	0%
Singh et al. (2020), Tsakalidis et al. (2015) and Brito and Adeodato (2020)	Linear regression	100%
Tsakalidis et al. (2015)	Sequential minimal optimization for regression	–
Heredia et al. (2017), Shaban et al. (2017), Heredia et al. (2018)	Convolutional neural networks (CNN)	50%
Bilal et al. (2018), Lopardo and Brambilla (2018)	Recurrent neural networks (RNN)	50%
Kalampokis et al. (2017)	DynamicLMC	–
Kassraie et al. (2017) and Tsakalidis et al. (2015)	Gaussian Process regression model	0%
Praciano et al. (2018)	Logistic regression	100%
Praciano et al. (2018), Joseph (2019), Singh et al. (2020), Bachhuber et al. (2016)	Decision tree	33.33%
Singh et al. (2020)	K-Nearest Neighbors	100%
Wang and Gan (2017) and Ibrahim et al. (2015)	Not informed	–
Awais et al. (2019)	Bayesian optimization model	0%
Bansal and Srivastava (2018)	Bitern Topic Model (BTM)	100%
Tung et al. (2016), Bansal and Srivastava (2018), Castro and Vaca (2017), Fano and Slanzi (2017)	Latent Dirichlet Allocation (LDA)	75%



**Fig. 8** Papers by approach

The column *paper* indicates the paper in which the approach was presented; the column *vol.* informs if the approach uses at least one method based only on post counting (i.e., volume-based, as explained in Sect. 2); the column *sent.* is checked when the approach presents at least one method based on sentiment analysis; the column *source* refers to the data source from where opinions were collected (e.g., Twitter or Facebook); the column *alig.* informs if the paper predicts the political alignment of tweets/users to forecast election outcomes; the column *ev.* refers to the ones that

use strategies related to event detection; the column *pop.* is checked when the paper proposes a means to calculate the candidate popularity; Finally, the column *success* indicates if the authors of the paper state that their approach predicted the election winner correctly. This field can be filled with four different labels:

- *Partial*—the authors state that their methods achieve success in at least one of their experiments but not in all experiments;
- *✓*—the authors state that their results predicted the election winner in all their experiments;
- *N/A*—the authors did not explicitly inform if their approach predicted the election winner;
- *No*—the authors state that their approach has failed to predict the election winner correctly.

Cases of partial success are summarized as follows:

- *Papers that present more than one election prediction method:* this is the case of works in Heredia et al. (2017), Bansal and Srivastava (2018), and Hwang (2019), which belong to the Counting-Based Approach and achieved

**Table 16** General characteristics of the surveyed papers

Ap.	Paper	Vol.	Sent.	Alig.	Ev.	Pop.	Source	Success
1	Maldonado and Sierra (2015)		✓				Twitter	✓
1	Almeida et al. (2015)		✓				Twitter	Partial
1	Dwi Prasetyo and Hauff (2015)		✓				Twitter	✓
1	Ibrahim et al. (2015)		✓				Twitter	✓
1	Khatua et al. (2015)		✓				Twitter	✓
1	Srivastava et al. (2015)		✓				Twitter	✓
1	Burnap et al. (2016)		✓				Twitter	✓
1	Jose and Chooralil (2016)		✓				Twitter	N/A
1	Sharma and Moh (2016)		✓				Twitter	Partial
1	Wicaksono et al. (2016)		✓				Twitter	✓
1	Sanders et al. (2016)	✓					Twitter	✓
1	Vepsäläinen et al. (2017)	✓					Facebook	N/A
1	Heredia et al. (2017)	✓	✓				Twitter	Partial
1	Rosseti et al. (2017)	✓					Twitter	No
1	Ramzan et al. (2017)		✓				Twitter	✓
1	Hinch (2017)	✓					Twitter	No
1	Singh et al. (2017)		✓				Twitter	✓
1	Praciano et al. (2018)	✓	✓				Twitter	Partial
1	Heredia et al. (2018)	✓	✓				Twitter	Partial
1	Bilal et al. (2018)	✓	✓				Twitter	✓
1	Naiknaware and Kawathekar (2018)	✓	✓				Twitter	N/A
1	Bansal and Srivastava (2018)	✓	✓				Twitter	Partial
1	Budiharto and Meiliana (2018)	✓	✓				Twitter	✓
1	Bansal and Srivastava (2019)	✓	✓				Twitter	✓
1	Hwang (2019)	✓	✓				Reddit	Partial
1	Kristiyanti et al. (2019)	✓	✓				Twitter	No
1	Singh et al. (2020)	✓	✓				Twitter	✓
1	Sanders and van den Bosch (2020)	✓	✓				Twitter	✓
2	Bachhuber et al. (2016)			✓			Twitter	No
2	Castro and Vaca (2017)			✓			Twitter	✓
2	Bastos and Mercea (2018)			✓			Twitter	No
2	Campanale and Caldarola (2018)			✓			Twitter	N/A
2	Lopardo and Brambilla (2018)			✓			Twitter	Partial
3	Unankard et al. (2014)		✓		✓		Twitter	✓
3	Tung et al. (2016)				✓		BSS	Partial
3	Shaban et al. (2017)	✓	✓		✓		Twitter	N/A
4	You et al. (2015)		✓			✓	Flickr	✓
4	Kagan et al. (2015)		✓			✓	Twitter	✓
4	Dokoohaki et al. (2015)					✓	Twitter	N/A
4	Wang and Lei (2016)	✓	✓			✓	Taiwan forum	✓
4	Xie et al. (2016)	✓				✓	Twitter, Facebook, Google, webpages	✓
4	Wang and Gan (2017)		✓			✓	Twitter	✓
4	Wang and Gan (2018)	✓	✓			✓	Twitter	N/A
4	Joseph (2019)	✓	✓			✓	Twitter	N/A
5	Kalampokis et al. (2017)	✓	✓				Twitter	N/A
5	Tsakalidis et al. (2015)		✓				Twitter	N/A
5	White (2016)	✓	✓				Twitter	Partial
5	Kassraie et al. (2017)		✓				Twitter	No
5	Fano and Slanzi (2017)		✓				Twitter	✓
5	Ajito et al. (2017)						Twitter, blogs	Partial

**Table 16** (continued)

Ap.	Paper	Vol.	Sent.	Alig.	Ev.	Pop.	Source	Success
5	Huang (2017)		✓			✓	Facebook, e-news, magazines	✓
5	Awais et al. (2019)		✓			✓	Twitter	Partial
5	Brito and Adeodato (2020)						Twitter, Facebook, Instagram, past elections, traditional polls	✓

success for the sentiment-based method and failed when using the volume-based method; the work presented in Ajito et al. (2017), which belongs to the Other Works section and achieved success when analyzed data from blogs and failed when analyzed data from Twitter;

- *Papers that only achieved success for some cities, states, districts or seats:* this is the case of the following Counting-Based approaches: (Almeida et al. 2015; Praciano et al. 2018; Heredia et al. 2018); the Event Detection Approach described in Tung et al. (2016); the Political Alignment Approach presented in Lopardo and Brambilla (2018); and the following works presented in the Other Works section: (White 2016; Awais et al. 2019).
- *Papers with more than one sentiment analysis method:* this is the case of the Counting-Based Approach (Sharma and Moh 2016) that failed when adopted dictionaries to predict sentiment and achieved success when adopted machine learning methods to infer sentiment.

We have observed that although several approaches state that they achieved success, they do not consider real percentages as, most of them, assume that the candidate associated with the higher number of (positive) instances is the election winner. Also, while some approaches compare their results with the real outcomes, other ones compare their results only with traditional polls. Another point that we can see from Tables 15 and 16 is that although some papers adopted more recent algorithms such as the ones based on some deep learning strategy, they failed to predict the winner or only achieved a partial success. This is the case of approaches (Heredia et al. 2017, 2018; Bastos and Mercea 2018), for example. On the other hand, we observed that approaches that adopted traditional machine learning algorithms were able to achieve a correct prediction in many cases, as is the case of Srivastava et al. (2015), Sharma and Moh (2016), Castro and Vaca (2017), Almeida et al. (2015), Sharma and Moh (2016), Dwi Prasetyo and Hauff (2015), and Wicaksono et al. (2016), for example. The amount of collected data is not directly related to success. We can see this by looking at Tables 16 and 7, where the approaches in Bastos and Mercea (2018) and Rosseti et al. (2017) failed to predict the election winner even though they were in the group that collected more tweets (more than 1,000,000 tweets). In the opposite side, papers in the group that collected fewer

tweets (less than 100,000 instances) were able in some cases to achieve success, namely Ramzan et al. (2017) and Maldonado and Sierra (2015). Another point that we can notice looking at Table 16 is that most of the works that achieved success predicted the opinions sentiment in their analysis, independently of the approach adopted. For instance, even though (Maldonado and Sierra 2015; Castro and Vaca 2017; Unankard et al. 2014; Wang and Lei 2016; Fano and Slanzi 2017) adopted different approaches (Counting-Based Approach, Political Alignment Approach, Event Detection Approach, Popularity-Based Approach, and Other Works, respectively), all of them combine their approaches with sentiment analysis. In addition, we also can observe that most of approaches that succeed (22 out of 25) use Twitter as source of opinions.

## 5 Discussions, limitations and challenges for future research

In this section, we resume the answer for the research questions Q1, Q2, Q3 presented in Sect. 3. We firstly present our categorization about the main election forecasting approaches using social media found for answering Q1). After that, we identify gaps and limitations in the current literature for answering Q2, both from opinion mining point of view, discussing limitations in regard to the proposals presented in the literature, and from tradition election polls point of view, discussing how opinion mining in social media can help leverage their results. Finally, we point out directions for future research, mainly from the machine learning and artificial intelligence point of view, for answering Q3.

### 5.1 Main election forecasting approaches

One of the goals of this research was to identify the main approaches for forecasting elections using social media. We observed that, in general, the surveyed approaches can be categorized into four main groups: (i) counting-based approach (Ramzan et al. 2017; Heredia et al. 2017; Sharma and Moh 2016; Jose and Chooralil 2016; Dwi Prasetyo and Hauff 2015; Burnap et al. 2016; Srivastava et al. 2015;

Wicaksono et al. 2016; Khatua et al. 2015; Maldonado and Sierra 2015; Almeida et al. 2015; Heredia et al. 2018; Ibrahim et al. 2015; Vepsäläinen et al. 2017; Rosseti et al. 2017; Praciano et al. 2018; Singh et al. 2020; Kristiyanti et al. 2019; Naiknaware and Kawathekar 2018; Bilal et al. 2018; Hinch 2017; Singh et al. 2017; Sanders et al. 2016; Bansal and Srivastava 2018; Sanders and van den Bosch 2020; Bansal and Srivastava 2019; Budiharto and Meiliana 2018; Hwang 2019)—this is the most simple approach, in which papers basically sum mentions to a specific party/candidate (volume-based) or sum the occurrence of positive opinions that mention a given party/candidate (sentiment-based) to predict election outcomes; (ii) political alignment approach (Bachhuber et al. 2016; Castro and Vaca 2017; Bastos and Mercea 2018; Campanale and Caldarola 2018; Lopardo and Brambilla 2018)—papers that try to predict the political alignment/leaning of the users to forecast election outcomes; (iii) event detection approach (Tung et al. 2016; Unankard et al. 2014; Shaban et al. 2017)—papers that relate the victory of a candidate/party to the occurrence of political events and predict outcomes based on that; (iv) Popularity-based approach (You et al. 2015; Wang and Gan 2017; Xie et al. 2016; Dokoochaki et al. 2015; Kagan et al. 2015; Wang and Lei 2016; Joseph 2019)—papers that propose to use a formula to infer candidates popularity and assume that the most popular candidate will win the election.

In what follows, we present a discussion about the surveyed papers according to their category. We mention some papers of each approach as example, highlighting some of their characteristics and emphasizing the ones that present particular aspects that differ them from other ones in the same category. In addition to the four approach categories identified in the literature, we create a category called *Other Works* (Kassraie et al. 2017; Ajito et al. 2017; Fano and Slanzi 2017; White 2016; Kalampokis et al. 2017; Tsakalidis et al. 2015; Brito and Adeodato 2020; Awais et al. 2019; Huang 2017) to group papers that do not fit in any of the identified categories.

### 5.1.1 Counting-based approach

We verified that works that *only* consider volume were not successful in most of the cases. Heredia et al. (2017), Khatua et al. (2015), Bansal and Srivastava (2018), Hwang (2019) and Heredia et al. (2018) tested both methods—sentiment- and volume-based. While Heredia et al. (2017), Bansal and Srivastava (2018), and Hwang (2019) only achieved good results using the sentiment-based strategy, Khatua et al. (2015) achieved the expected election winner with both methods. Heredia et al. (2018) concluded that the volume-based result was equivalent to the sentiment-based result on the national level. However, in their experiments, the sentiment-based method outperformed the volume-based

when it comes to state level elections. Sharma and Moh (2016) tested two methods using sentiment-based strategy, the first one uses machine learning algorithms and the other one is based only on dictionaries. In their case, they only predicted the correct winner in the experiments that use machine learning algorithms. Other approaches such as the one presented by Kristiyanti et al. (2019) did not predict the correct election winner using the sentiment-based strategy. Vepsäläinen et al. (2017) adopted a counting approach based on the number of Facebook likes and concluded that election prediction based on Facebook is not accurate when compared to traditional polls.

Some papers adopted counting-based approaches that are more elaborated. Dwi Prasetyo and Hauff (2015) go further by filtering non-personal accounts, slacktivists and spam users. In this context, Ibrahim et al. (2015) built a classifier to detect buzzer accounts aiming at removing data noise. Almeida et al. (2015) also removed some spam and news accounts and considered demographic information. Dwi Prasetyo and Hauff (2015) used both demographic and geolocation information. Srivastava et al. (2015) and Khatua et al. (2015) tried to predicted not only vote share but also seat share. Srivastava et al. (2015) and Singh et al. (2020) tried to find out the number of seats. Wicaksono et al. (2016) and Singh et al. (2020) presented a subtle difference in relation to the other sentiment-based approaches since this approach (Wicaksono et al. 2016) also takes into account the number of tweets with negative sentiment. While Wicaksono et al. (2016) assume that a negative opinion can be interpreted as a vote to the opposition, Singh et al. (2020) presented a formula to compute the actual sentiment score that depends on the negative score. Naiknaware and Kawathekar (2018) adopted a counting-based approach to find out the most supported agenda and from that predict the election winner. Finally, Bansal and Srivastava (2018) find out tweet topics using word co-occurrences and infer the sentiment of the tweet based on the sentiments of the topics.

### 5.1.2 Political alignment approach

From the works that classify tweets according to the political alignment (e.g., as “republican” or “democrat” (Lopardo and Brambilla 2018)), the success rate was variable. While Campanale and Caldarola (2018) argued that they achieved promising results, Castro and Vaca (2017) predicted the correct winner, Bastos and Mercea (2018) and Bachhuber et al. (2016) failed to predict the right result, and Lopardo and Brambilla (2018) presented the correct prediction only for some districts.

Some of the papers were concerned with other aspects in their analysis. For example, Castro and Vaca (2017) proposed to compute a reputation score based on the number of



friends and followers that a Twitter account has to discard bot accounts. Castro and Vaca (2017) and Bastos and Mercea (2018) take into account geolocation information.

### 5.1.3 Event detection approach

Tung et al. (2016) assume that events can have positive or negative impact on people opinions in relation to a given candidate. They try to detect the occurrence of events that are related to the elections based on terms that appear in users comments on an online forum. By using rules, they determine the winner party based on events occurrence. With this approach, they predicted the election winner correctly for some cities. On the other hand, Shaban et al. (2017) and Unankard et al. (2014) use event detection approaches that were successful to predict the election winner. Unankard et al. (2014) present a more elaborated approach, by clustering tweets that belong to the same event/topic based on their terms. They use part-of-speech tagging and named entity recognition to determine events location. Additionally, the work by Unankard et al. (2014) also predicted tweets sentiment and proposed some rules to detect sarcasm constructions.

### 5.1.4 Popularity-based approach

You et al. (2015) achieved a successful result using a strategy that try to predict the candidates popularity by combining textual and image features using data from Flickr. Wang and Gan (2017) and Xie et al. (2016) were also successful to predict the correct winner. Different from the other approaches that were grouped into this category, Xie et al. (2016) use demographic information and considered many sources of information to infer candidates popularity (Twitter, Facebook, Google, and candidates' campaign websites and offline data from pollsters). Dokoohaki et al. (2015) and Kagan et al. (2015) presented approaches where the candidate popularity was computed based on an analysis of graphs of user interactions in Twitter, achieving good results. Wang and Lei (2016) also argued that they predicted the election winner correctly, by using an approach that takes into account rating records of candidate related articles in a popular Taiwan forum. Joseph (2019) proposed a popularity formula that is based on sentiment score of positive and neutral tweets that achieved the correct result. Wang and Gan (2018) predicted the correct winner in one of their experiments, which computes candidate popularity based on the score of keywords related to him.

We noticed that most of the works that achieved successful predictions adopted a sentiment analysis step and used Twitter as source of opinions, independently of the approach adopted. For this reason, we believe that this may indicate that Twitter is a promising source for collecting

electoral opinions and that sentiment analysis should be considered for those who want to achieve better electoral predictions using social media. An example of this are the papers: (Maldonado and Sierra 2015; Castro and Vaca 2017; Unankard et al. 2014; Wang and Lei 2016) and (Fano and Slanzi 2017), which adopted different approaches (counting-based approach, political alignment approach, event detection approach, popularity-based approach, and other works, respectively). Also, most of the successful works have attempted to predict presidential election results, which leads us to believe that social media election surveys are more suitable for analyzing elections at national level.

## 5.2 Limitations and challenges for data science

There are many aspects that can lead to wrong predictions in traditional polls. Sturgis et al. (2018) and Zeedan (2019) pointed out that last minute changes, i.e., a shift in vote share toward one of the parties between the final polls and Election Day, may be one of the reasons for wrong predictions in traditional polls. Breur (2016) highlights that fake news and social media bots had a high influence on voters opinions in the 2016 US Presidential elections, factor that could be responsible for vote changes in a short period. According to Michael Bruter (2017) (Castelvecchi 2017), a political scientist at the London School of Economics, another reason for wrong predictions is the fact that some people only make up their minds on the eve of the election. Zeedan (2019) argues that wrong predictions may also occur when pollsters fail to achieve a representative sample, due to the lack of accurate phone databases or when pollsters assume that people who did not vote in past elections will not vote in the next elections, for example. In this way, opinion mining and social media data analysis can be a helpful tool for leveraging the prediction power of traditional tools for (i) detecting fake news and social media bots; (ii) detecting opinion changing of the people regarding the candidates over time; and (iii) identification of different types of voters considering an analysis of their behavior and profile, increasing the representativeness of the population in the polls. In the last case, tools for supporting the identification of voters profile may be adopted.

On the other hand, although social media has emerged as a promising way of collecting opinions in real-time, there are yet many limitations when social media data is used to predict election outcomes from computer science point of view:

- *Methods cannot be generalized:* Usually, the methods found in the literature for election forecast using social media only consider specific elections, implying that the results cannot be general enough to contemplate other elections;

- *Non-availability of datasets*: One of the gaps found in this domain is that the electoral datasets are not freely available for the community that work on this topic. It is not possible to evaluate the success of existing forecasting methods without analyzing their datasets. We cannot guarantee, for example, if they were successful because of the election (i.e., the case in which the election is not a close dispute) or if they were successful due to the effectiveness of the adopted methodology. We believe that this can be the reason of some similar forecasting strategies achieving the correct prediction in some papers (that refer to a given election) and the wrong in other ones. In order to address this issue, a temporal analysis could be conducted evaluating data from time to time in different time periods that preceded the election to check if the predicted election winner changes over time. Also, we cannot compare existing approaches by predictive accuracy since they refer to different datasets/elections;
- *Filtering potential users*: Data from social media can be posted by non-person users, such as organizations. Few approaches discard this kind of post. Additionally, the majority of approaches consider in their calculus different posts by the same user. This can impact the final results since each person only can vote one time. Another problem is that most of these approaches also did not analyze if the social media user account belongs to a person that is permitted to vote in the given election or even if it belongs to a real person, i.e., they did not discard fake accounts and accounts that belong to non-voters. In this context, data collected from social media can be posted by bots (spam). Therefore, the nature of data collected from the Web can have many biases, reducing trust and the credibility of the results. Furthermore, the general profile of people that use social media is different from the voters, mainly in developing and non-developing countries, i.e., the majority of Twitter users are young men that live in urban areas as pointed out in Dwi Prasetyo and Hauff (2015). Taking into account these factors, sometimes a big amount of data cannot reflect a statistically representative sample of the general population;
- *Sentiment analysis challenges: Data Labeling*: We noticed that many approaches that forecast the election outcomes based on sentiment analysis rely on straightforward methods for labeling the sentiment of the sentences (such as emoticons or dictionaries) (see Sect. 4.3 and Fig. 6), ignoring that predictions may be misleading due to *the difference between domains*. This is because the polarity of a word depends on the context that it is inserted. For instance, the word *scary* can express a negative sentiment when extracted from posts related to general contexts and positive sentiment when extracted from opinions about horror movies. In addition, terms that denote sentence sentiment can vary according to the domain, i.e., although the word *cheap* indicates a positive sentiment in the product reviews domain, this word does not denote a positive sentiment in tweets talking about a given political candidate; *Sarcasm and Irony*: Another challenge to infer sentiment polarity of texts is that ironic and sarcastic posts are prevalent on social media. In this way, a text thanking a person, for example, can be expressing the contrary opinion. Although was observed by recent approaches social media texts related to elections are full of ironic content (Duarte et al. 2019), only one work analyzed in this literature survey takes into account this issue. However, even in such a case only a simple mechanism is presented to deal with sarcasm/irony;
- *Absence of a methodological pattern*: There is not a default methodology to predict elections based on data from social media, i.e., approaches use different steps to collect data and estimate the prediction. We have observed that each research surveyed considers different periods to begin/end the data collection. Additionally, each one of them collect data containing different kinds of terms (for instance, candidates' names, parties' names, campaign slogans, and so on) and considering the different quantity of posts (see Sect. 4.1);
- *Accuracy of the polls based on social media*: In general, predictions based on polls that use social media have lower accuracy than predictions based on traditional polls, i.e., most of times traditional polls present results closer to the actual results. However, data analysis and opinion mining tools on social media in election scenarios can be very important for improving traditional polls results;
- *Absence of patterns for evaluating the predictions*: There is not a consensus in the literature about how to evaluate the predictions. While some papers compare their prediction with the real predictions (post hoc analysis), other ones compare their result with the result of predictions based on traditional polls. Also, the majority of approaches that argue that they were successful only take into account the absolute election winner (and not vote share);
- *Post hoc analysis*: Several approaches present a post hoc analysis, analyzing social media data, and calculating the prediction after the occurrence of the real election. According to Gayo-Avello (2012), this cannot be considered a *prediction* at all. So, developing tools for tracking social media users behavior regarding their candidates is very important in this scenario;
- *Elector behavior*: as pointed out by Wang and Gan (2017), the behavior of the elector can affect the accuracy of the predictions methods based on social media. This is because while most supporters of a given candi-

date  $A$  may not attack its adversaries on social media, the supporters of a candidate  $B$  may usually attack the other ones, posting a huge amount of data;

- *Annotator bias*: Supervised machine learning techniques for sentiment analysis/opinion mining rely on labeled datasets that can be annotated by a small set of persons that also do not reflect the characteristics of the electorate. This fact may affect results on classification tasks.

### 5.3 Open issues and future research from the AI point of view

In what follows, we highlight some lines for future research in opinion mining for elections outcomes predictions:

- *Opinion mining using multimodal data*: Only one of the papers surveyed considered images shared on social media in order to predict election results (You et al. 2015). This field could be better explored by future related lines of research since many social media posts contain images and not only text.
- *Data streaming mining*: We have identified a lack of approaches that deal with election forecast using data stream mining methods. These methods are interesting as they adapt the machine learning model over time (Bifet et al. 2018). Moreover, there are also methods that allow to identify concept drift, which could be interesting in electoral scenario. On the other hand, one main challenge for this is acquiring labeled data. In this way, unsupervised or semi-supervised data stream mining methods could be explored;
- *Active learning*: We have noticed that none of the papers in our survey adopt active learning methods (Tong and Koller 2001). These methods include the human in the machine learning loop. Basically, they select the most representative instances to be labeled by humans to deal with the lack of labeled data in diverse domains. These methods can be investigated in future researches to cope with the problem of short period for labeling domain specific data, helping to improve election outcomes prediction;
- *Domain adaptation and transfer learning*: We did not find in the literature papers that predict elections using transfer learning and domain adaptation strategies (Pan et al. 2010). Pre-trained word embedding techniques could be better explored in order to improve the accuracy of the results. Additionally, recent lines of research based on language modeling, such as ULMFit (Howard and Ruder 2018), ELMo (Peters et al. 2018) and BERT (Devlin et al. 2019), could also be investigated. These methods take advantage of the ability of language modeling networks to representation and semantics and fine-tuning them to perform classification tasks. Also,

as a good practice, the community that work on topics related to electoral domain problems could concentrate some efforts to make available existing labeled domain datasets and existing machine learning models in order to facilitate forecasting tasks of future elections as well to enable additional analyses of existing ones. Another important aspect of this line of research is verifying the possibility of transfer learning considering datasets of different languages (datasets from previous elections in different languages could be used) or different domains, as well as proposing new mechanisms to search which datasets are more suitable to this task. This is an important line of research as the number of datasets available on Internet for opinion mining grows and the number of experiments needed to choose the appropriated classifier can also exponentially grow.

- *Gamification*: Given the importance of having labeled data in the electoral domain to build reliable classifiers, we believe that gamification labeling strategies (Öhman et al. 2018) can be explored to motivate manual labeling of social media electoral opinions.

## 6 Conclusions

Election polls are fundamental components in democratic societies and can be used with many purposes, e.g., measure vote intention, adjust electoral campaigns, and predict favoritism or rejection. In the electoral scenario, forecasting methods must ideally present not only accurate predictions but also must be able to predict in a timely and cost-effective way (Wang et al. 2015). By using social media such as Twitter, it is possible to monitor users' opinions continuously. In this manuscript, we surveyed approaches that use data posted in social media to infer election outcomes. Those approaches have been proposed as alternatives to traditional election polls. We conducted a systematic review and organize the literature along several aspects such as the social media-based mechanisms used to predict political outcomes, the quantity of collected data, the specific social media used, the collection period, the algorithms and approaches adopted, and so on. From this analysis, we have identified the main approaches that have been proposed to deal with this problem, the main tasks performed by those approaches, and point out limitations and lines for further research on this topic. One of the gaps identified is the non-availability of datasets and machine learning models related to the electoral domain. From the trustworthy AI perspective, this issue is problematic since data, processes and algorithms must be shared to enable that in-depth analyses are carried out (Janssen et al. 2020).

An essential difference between traditional and social media-based methods to election forecast is that while

people post thoughts and opinions on the Internet spontaneously, without (any) external pressure and in a free format (Caldarelli et al. 2014), traditional polls rely on intermediaries that ask people concrete questions. Since people may not feel comfortable to express their real opinions in the presence of an intermediary, it could affect their answers. Additionally, traditional polls can have high non-response rates (Vepsäläinen et al. 2017), and interviewers might alter the sense of questions and influence answers, for instance, according to their discourse tone (Maldonado and Sierra 2015). Another positive factor of social media-based approaches relies on the fact that it can collect real-time data automatically, with lower cost. This factor is specially important since many voters tend to decide their votes in the last 24 h before the Election Day (Mitchell 1992; Castelvichi 2017). Furthermore, *last minute changes*—which occur when people change their minds on the eve of elections—are also pointed out as one of the possible reasons for wrong predictions in traditional election forecasting methods (Sturgis et al. 2018; Zeedan 2019).

In a general summary, polls based on information posted on social media have a set of advantages when compared with traditional ones: they can capture data automatically, in real-time, and at a lower cost. On the other hand, the outcomes of polls that use social media can be affected by noisy data (You et al. 2015) and for age-bias due to the difference in the age groups between people that use social media and people target by polling agencies (Xie et al. 2016). For this reason, it is important to interpret and analyze social media data regarding elections instead of simply create a direct correspondence between each social media post related to a candidate and an election vote.

We consider that the list of limitations and lines for future research presented in Sect. 5 should be taken into account by the ones that desire to gather political opinions from social media to analyze electoral aspects and the ones that want to propose new approaches to forecast election results based on social media. Finally, we concluded with this survey that social media can be adopted as *thermometers* on the political campaigns during an election but they are still not enough to replace traditional methods entirely, where samples are more representative. We believe that social media analysis can be very interesting to be explored in scenarios in which very disputed elections take place, and last minute voting changes have the power to affect the final election outcome. Therefore, approaches that mix traditional election forecasting with social media analysis could be explored to try to get better results.

As a future work, we intend to investigate some of the topics mentioned in Sect. 5.3, aiming to build a robust and trusted information sharing framework for election analysis based on social media. We will begin by exploring transfer learning and active learning strategies to improve opinions

analysis regarding the electoral scenario, where we have complex opinions to be analyzed and a short period for labeling data. Also, we want to provide for the community a manually labeled dataset of tweets related to the electoral domain, aiming that it can be used to assist analysis of future elections.

**Acknowledgements** We would like to thank the Brazilian Research CNPq APQ Universal (Grant 421608/2018-8), CNPq Research Grant 311275/2020-6, FAPERJ Research grant E26/202.914/2019 (247109), LATAM Microsoft Research Grant for the financial support and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brazil (CAPES)—Finance Code 001, for the scholarship granted to the first author.

**Author Contributions** All authors contributed to the study conception and design. The literature search was performed by Jéssica S. Santos, and data analysis was performed by all authors. The first draft of the manuscript was written by Jéssica S. Santos, and Flavia Bernardini and Aline Paes commented on previous versions of the manuscript and critically revised the work. All authors read and approved the final manuscript.

**Funding** This research was supported by the Brazilian Research CNPq APQ Universal (Grant 421608/2018-8), CNPq Research Grant 311275/2020-6, FAPERJ Research Grant E26/202.914/2019 (247109), Microsoft Research Grant and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brazil (CAPES).

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

- Ajito M, Kawahata Y, Ishii A (2017) Analysis of national election using mathematical model of hit phenomenon. In: 2017 IEEE international conference on big data (Big Data). IEEE, pp 4722–4724
- Almeida JM, Pappa GL et al (2015) Twitter population sample bias and its impact on predictive outcomes: a case study on elections. In: Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining 2015. ACM, pp 1254–1261
- Asur S, Huberman BA (2010) Predicting the future with social media. In: Proceedings of the 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology—volume 01. IEEE Computer Society, pp 492–499
- Awais M, Hassan S-U Ahmed A (2019) Leveraging big data for politics: predicting general election of Pakistan using a novel rigged model. *J Ambient Intell Human Comput*, pp 1–9
- Bachhuber J, Koppeel C, Morina J, Rejström K, Steinschulte D (2016) Us election prediction: a linguistic analysis of us twitter users. In: Designing networks for innovation and improvisation. Springer, pp 55–63
- Bansal B, Srivastava S (2018) On predicting elections with hybrid topic based sentiment analysis of tweets. *Proc Comput Sci* 135:346–353
- Bansal B, Srivastava S (2019) Lexicon-based twitter sentiment analysis for vote share prediction using emoji and n-gram features. *Int J Web Based Commun* 15(1):85–99

- Bastos M, Mercea D (2018) Parametrizing Brexit: mapping twitter political space to parliamentary constituencies. *Inform Commun Soc* 21(7):921–939
- Bifet A, Gavaldà R, Holmes G, Pfahringer B (2018) *Machine learning for data streams with practical examples in MOA*. MIT Press. <https://moa.cms.waikato.ac.nz/book/>
- Bilal M, Asif S, Yousuf S, Afzal U, (2018) 2018 Pakistan general election: understanding the predictive power of social media. In: 2018 12th international conference on mathematics, actuarial science, computer science and statistics (MACS). IEEE, pp 1–6
- Bovet A, Morone F, Makse HA (2018) Validation of twitter opinion trends with national polling aggregates: Hillary clinton vs donald trump. *Sci Rep* 8(1):8673
- Breuer T (2016) US elections: how could predictions be so wrong?
- Brito KdS, Adeodato PJL (2020) Predicting Brazilian and us elections with machine learning and social media data. In: 2020 international joint conference on neural networks (IJCNN). IEEE, pp 1–8
- Budiharto W, Meiliana M (2018) Prediction and analysis of Indonesia presidential election from twitter using sentiment analysis. *J Big data* 5(1):1–10
- Burnap P, Gibson R, Sloan L, Southern R, Williams M (2016) 140 characters to victory?: Using twitter to predict the UK 2015 general election. *Electoral Stud* 41:230–233
- Calais Guerra PH, Veloso A, Meira W Jr, Almeida V (2011) From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 150–158
- <>e95809Caldarelli G, Chessa A, Pammolli F, Pompa G, Puliga M, Riccaboni M, Riotta G (2014) A multi-level geographical study of Italian political elections from twitter data. *PLoS ONE* 9(5):e95809
- Campanale M, Caldarola EG (2018) Revealing political sentiment with twitter: the case study of the 2016 Italian constitutional referendum. In: 2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). IEEE, pp 861–868
- Castelvecchi D (2017) Why the polls got the UK election wrong? *Nat News*
- Castro R, Vaca C (2017) National leaders' twitter speech to infer political leaning and election results in 2015 Venezuelan parliamentary elections. In: 2017 IEEE international conference on data mining workshops (ICDMW). IEEE, pp 866–871
- Chauhan P, Sharma N, Sikkka G (2021) The emergence of social media data and sentiment analysis in election prediction. *J Ambient Intell Humaniz Comput* 12(2):2601–2627
- Cornfield M (2008) Yes, it did make a difference. *Media Polit*
- Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp 4171–4186. <https://aclanthology.org/N19-1423>
- Di Giovanni M, Brambilla M, Ceri S, Daniel F, Ramponi G (2018) Content-based classification of political inclinations of twitter users. In: 2018 IEEE international conference on big data (Big Data). IEEE, pp 4321–4327
- Dokoohaki N, Zikou F, Gillblad D, Matskin M (2015) Predicting Swedish elections with twitter: a case for stochastic link structure analysis. In: 2015 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). IEEE, pp 1269–1276
- dos Santos Brito K, Silva Filho RLC, Adeodato PJL (2021) A systematic review of predicting elections based on social media data: research challenges and future directions. *IEEE Trans Comput Soc Syst*
- dos Santos JS, Paes A, Bernardini F (2019) Combining labeled datasets for sentiment analysis from different domains based on dataset similarity to predict electors sentiment. In: 2019 8th Brazilian conference on intelligent systems (BRACIS). IEEE, pp 455–460
- Duarte L, Macedo L, Oliveira HG (2019) Exploring emojis for emotion recognition in Portuguese text. In: EPIA conference on artificial intelligence. Springer, pp 719–730
- Dwi Prasetyo N, Hauff C (2015) Twitter-based election prediction in the developing world. In: Proceedings of the 26th ACM conference on hypertext and social media. ACM, pp 149–158
- Esuli A, Sebastiani F (2007) Sentiwordnet: a high-coverage lexical resource for opinion mining. *Evaluation* 17:1–26
- Fano S, Slanzi D (2017) Using twitter data to monitor political campaigns and predict election results. In: International conference on practical applications of agents and multi-agent systems. Springer, pp 191–197
- Forsythe R, Myerson RB, Rietz TA, Weber RJ (1993) An experiment on coordination in multi-candidate elections: the importance of polls and election histories. *Soc Choice Welfare* 10(3):223–247
- Gao L, Kuppersmith A, Huang R (2017) Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In: Proceedings of the eighth international joint conference on natural language processing (volume 1: long papers). Asian federation of natural language processing, Taipei, Taiwan, pp 774–782. <https://aclanthology.org/I17-1078>
- Garcia ACB, Silva W, Correia L (2018) The prednews forecasting model. In: Proceedings of the 19th annual international conference on digital government research: governance in the data age, pp 1–6
- Gayo-Avello D (2011) Don't turn social media into another 'literary digest' poll. *Commun ACM* 54(10):121–128
- Gayo-Avello D (2012) No, you cannot predict elections with twitter. *IEEE Internet Comput* 16(6):91–94
- Gelman A, King G (1993) Why are American presidential election campaign polls so variable when votes are so predictable? *Br J Polit Sci* 23(4):409–451
- Graefe A (2014) Accuracy of vote expectation surveys in forecasting elections. *Public Opin Q* 78(S1):204–232
- Heredia B, Prusa JD, Khoshgoftaar TM (2018) Social media for polling and predicting united states election outcome. *Soc Netw Anal Min* 8(1):48
- Heredia B, Prusa J, Khoshgoftaar T (2017) Exploring the effectiveness of twitter at polling the united states 2016 presidential election. In: 2017 IEEE 3rd international conference on collaboration and internet computing (CIC). IEEE, pp 283–290
- Hinch J (2017) #makeamericaspollsgreatagain: Evaluating twitter as a tool to predict election outcomes
- Howard J, Ruder S (2018) Universal language model fine-tuning for text classification. In: Proceedings of the 56th annual meeting of the association for computational linguistics, pp 328–339
- Huang J-Y (2017) Web mining for the mayoral election prediction in Taiwan. *Aslib J Inform Manag* 69(6):688–701
- Hwang B (2019) Reddit sentiment analysis to improve election predictions. In: International conference big data analytics. Data mining and computational intelligence, pp 204–208
- Ibrahim M, Abdillah O, Wicaksono AF, Adriani M (2015) Buzzer detection and sentiment analysis for predicting presidential election results in a twitter nation. In: 2015 IEEE international conference on data mining workshop (ICDMW). IEEE, pp 1348–1353

- Idan L, Feigenbaum J (2019) Show me your friends, and i will tell you whom you vote for: predicting voting behavior in social networks. In: 2019 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). IEEE, pp 816–824
- Janssen M, Brous P, Estevez E, Barbosa LS, Janowski T (2020) Data governance: organizing data for trustworthy artificial intelligence. *Govern Inform Q* 37(3):101493
- Jose R, Choorail VS (2016) Prediction of election result by enhanced sentiment analysis on twitter data using classifier ensemble approach. In: 2016 international conference on data mining and advanced computing (SAPIENCE). IEEE, pp 64–67
- Joseph FJJ (2019) Twitter based outcome predictions of 2019 Indian general elections using decision tree. In: 2019 4th international conference on information technology (InCIT). IEEE, pp 50–53
- Jungherr A, Jürgens P, Schoen H (2012) Why the pirate party won the German election of 2009 or the trouble with predictions: a response to tumasjan, a., sprenger, to, sander, pg, and welpel, im “predicting elections with twitter: what 140 characters reveal about political sentiment. *Soc Sci Comput Rev* 30(2):229–234
- Kagan V, Stevens A, Subrahmanian V (2015) Using twitter sentiment to forecast the 2013 Pakistani election and the 2014 Indian election. *IEEE Intell Syst* 30(1):2–5
- Kalampokis E, Karamanou A, Tambouris E, Tarabanis KA (2017) On predicting election results using twitter and linked open data: the case of the UK 2010 election. *J. UCS* 23(3):280–303
- Kassraie P, Modirshanechi A, Aghajan HK (2017) Election vote share prediction using a sentiment-based fusion of twitter data with google trends and online polls. In: *DATA*, pp 363–370
- Katz RS (1997) *Democracy and elections*. Oxford University Press on Demand
- Khatua A, Khatua A, Ghosh K, Chaki N (2015) Can# twitter\\_trends predict election results? evidence from 2014 Indian general election. In: 2015 48th Hawaii international conference on system sciences (HICSS). IEEE, pp 1676–1685
- Koli AM, Ahmed M (2019) Exploring the power of social media in election outcomes. *Int J Recent Technol Eng*
- Kristiyanti DA, Umam AH et al. (2019) Prediction of Indonesia presidential election results for the 2019–2024 period using twitter sentiment analysis. In: 2019 5th international conference on new media studies (CONMEDIA). IEEE, pp 36–42
- Li B, Guo D, Chang M, Li M, Bian A (2017) The prediction on the election of representatives. In: 2017 international conference on security, pattern analysis, and cybernetics (SPAC). IEEE, pp 329–334
- Liu B (2020) *Sentiment analysis: mining opinions, sentiments, and emotions, studies in natural language processing*, 2nd edn. Cambridge University Press, Cambridge
- Lopardo A, Brambilla M (2018) Analyzing and predicting the us mid-term elections on twitter with recurrent neural networks. In: 2018 IEEE international conference on big data (Big Data). IEEE, pp 5389–5391
- Mahendiran A, Wang W, Lira JAS, Huang B, Getoor L, Mares D, Ramakrishnan N (2014) Discovering evolving political vocabulary in social media. In: 2014 international conference on behavioral, economic, and socio-cultural computing (BESCom2014). IEEE, pp 1–7
- Maldonado M, Sierra V (2015) Can social media predict voter intention in elections? \\_x000d\\_ the case of the 2012 dominican republic presidential election
- Mitchell V-W (1992) Opinion polls: right or wrong? A lesson in social research. *Market Intell Plan* 10(9):4–9
- Naiknaware BR, Kawathekar SS (2018) Prediction of 2019 Indian election using sentiment analysis. In: 2018 2nd international conference on I-SMAC (IoT in social, mobile, analytics and cloud) (I-SMAC) I-SMAC (IoT in social, mobile, analytics and cloud) (I-SMAC). IEEE, pp 660–665
- Nguyen D, Trieschnigg D, Meder T (2014) Tweetgenie: Development, evaluation, and lessons learned. In: *Proceedings of COLING 2014, the 25th international conference on computational linguistics: system demonstrations*, pp 62–66
- Öhman E, Kajava K, Tiedemann J, Honkela T (2018) Creating a dataset for multilingual fine-grained emotion-detection using gamification-based annotation. In: *Proceedings of the 9th workshop on computational approaches to subjectivity. Sentiment and social media analysis*, pp 24–30
- Okeowo A (2016) *Hate on the rise after trump’s election*. The New Yorker
- Pan SJ, Yang Q et al (2010) A survey on transfer learning. *IEEE Trans Knowl Data Eng* 22(10):1345–1359
- Pennebaker JW, Francis ME, Booth RJ (2001) *Linguistic inquiry and word count: Liwc 2001*. Mahway: Lawrence Erlbaum Associates
- Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. In: *Proceedings of North American Association for Computational Linguistics (NAACL)*
- Praciano BJB, da Costa JPCL, Maranhão JPA, de Mendonça FLL, de Sousa Júnior RT, Pretz JB (2018) Spatio-temporal trend analysis of the Brazilian elections based on twitter data. In: 2018 IEEE international conference on data mining workshops (ICDMW). IEEE, pp 1355–1360
- Ramzan M, Mehta S, Annapoorna E (2017) Are tweets the real estimators of election results? In: 2017 tenth international conference on contemporary computing (IC3). IEEE, pp 1–4
- Rosseti I, Viterbo J et al. (2017) On tweets, retweets, hashtags and user profiles in the 2016 American presidential election scene. In: *Proceedings of the 18th annual international conference on digital government research*. ACM, pp 120–128
- Rothschild D, Malhotra N (2014) Are public opinion polls self-fulfilling prophecies? *Res Polit* 1(2):2053168014547667
- Sagiroglu S, Sinanc D (2013) Big data: a review. In: 2013 international conference on collaboration technologies and systems (CTS). IEEE, pp 42–47
- Sanders E, de Gier M, van den Bosch A (2016) Using demographics in predicting election results with twitter. In: *International conference on social informatics*. Springer, pp 259–268
- Sanders E, van den Bosch A (2020) Optimising twitter-based political election prediction with relevance and sentiment filters. In: *Proceedings of The 12th language resources and evaluation conference*, pp 6158–6165
- Santos JS, Bernardini F, Paes A (2021) Measuring the degree of divergence when labeling tweets in the electoral scenario. In: *Brazilian workshop on social network analysis and mining (BraSNAM 2021)*
- Shaban TA, Hexter L, Choi JD (2017) Event analysis on the 2016 us presidential election using social media. In: *International conference on social informatics*. Springer, pp 201–217
- Sharma P, Moh T-S (2016) Prediction of Indian election using sentiment analysis on Hindi twitter. In: 2016 IEEE international conference on big data (Big Data). IEEE, pp 1966–1971
- Singh P, Dwivedi YK, Kahlon KS, Pathania A, Sawhney RS (2020) Can twitter analytics predict election outcome? an insight from 2017 Punjab assembly elections. *Govern Inform Q*, p 101444
- Singh P, Sawhney RS (2018) Influence of twitter on prediction of election results. In: *Progress in advanced computing and intelligent engineering*. Springer, pp 665–673
- Singh P, Sawhney RS, Kahlon KS (2017) Predicting the outcome of Spanish general elections 2016 using twitter as a tool. In: *Advanced informatics for computing research*. Springer, pp 73–83

- Skoric MM, Liu J, Jaidka K (2020) Electoral and public opinion forecasts with social media data: a meta-analysis. *Information* 11(4):187
- Sokolova K, Perez C (2018) Elections and the twitter community: the case of right-wing and left-wing primaries for the 2017 French presidential election. In: 2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). IEEE, pp 1021–1026
- Srivastava R, Kumar H, Bhatia M, Jain S (2015) Analyzing Delhi assembly election 2015 using textual content of social network. In: Proceedings of the sixth international conference on computer and communication technology 2015. ACM, pp 78–85
- Sturgis P, Kuha J, Baker N, Callegaro M, Fisher S, Green J, Jennings W, Lauderdale BE, Smith P (2018) An assessment of the causes of the errors in the 2015 UK general election opinion polls. *J R Stat Soc Ser A Stat Soc* 181(3):757–781
- Taboada M (2016) Sentiment analysis: an overview from linguistics
- Taboada M, Brooke J, Tofiloski M, Voll K, Stede M (2011) Lexicon-based methods for sentiment analysis. *Comput Linguist* 37(2):267–307
- Tong S, Koller D (2001) Support vector machine active learning with applications to text classification. *J Mach Learn Res* 2(Nov):45–66
- Tsakalidis A, Papadopoulos S, Cristea AI, Kompatsiaris Y (2015) Predicting elections for multiple countries using twitter and polls. *IEEE Intell Syst* 30(2):10–17
- Tumasjan A, Sprenger TO, Sandner PG, Welpel IM (2010) Predicting elections with twitter: what 140 characters reveal about political sentiment. *Icwsn* 10(1):178–185
- Tung K-C, Wang ET, Chen AL (2016) Mining event sequences from social media for election prediction. In: Industrial conference on data mining. Springer, pp 266–281
- Unankard S, Li X, Sharaf M, Zhong J, Li X (2014) Predicting elections from social networks based on sub-event detection and sentiment analysis. In: International conference on web information systems engineering. Springer, pp 1–16
- Vania C, Ibrahim M, Adriani M (2014) Sentiment lexicon generation for an under-resourced language. *Int J Comput Linguist Appl* 5(1):59–72
- Vepsäläinen T, Li H, Suomi R (2017) Facebook likes and public opinion: predicting the 2015 Finnish parliamentary elections. *Govern Inform Q* 34(3):524–532
- Wang L, Gan JQ (2017) Prediction of the 2017 French election based on twitter data analysis. In: Computer science and electronic engineering (CEECE), 2017. IEEE, pp 89–93
- Wang L, Gan JQ (2018) Prediction of the 2017 French election based on twitter data analysis using term weighting. In: 2018 10th computer science and electronic engineering (CEECE). IEEE 231–235
- Wang M-H, Lei C-L (2016) Boosting election prediction accuracy by crowd wisdom on social forums. In: 2016 13th IEEE annual consumer communications and networking conference (CCNC). IEEE, pp 348–353
- Wang W, Rothschild D, Goel S, Gelman A (2015) Forecasting elections with non-representative polls. *Int J Forecast* 31(3):980–991
- White K (2016) Forecasting Canadian elections using twitter. In: Canadian conference on artificial intelligence. Springer, pp 186–191
- Wicaksono AJ, et al (2016) A proposed method for predicting us presidential election by analyzing sentiment in social media. In: 2016 2nd international conference on science in information technology (ICSITech). IEEE, pp 276–280
- Woolley SC (2016) Automating power: social bot interference in global politics. *First Monday* 21(4)
- Xie Z, Liu G, Wu J, Wang L, Liu C (2016) Wisdom of fusion: prediction of 2016 Taiwan election with heterogeneous big data. In: 2016 13th international conference on service systems and service management (ICSSSM). IEEE, pp 1–6
- You Q, Cao L, Cong Y, Zhang X, Luo J (2015) A multifaceted approach to social multimedia-based prediction of elections. *IEEE Trans Multimed* 17(12):2271–2280
- Zeedan R (2019) The 2016 US presidential elections: What went wrong in pre-election polls? Demographics help to explain. *J-Multidiscipl Sci J* 2(1):84–101

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations