PREDICTION OF USA NOVEMBER 2020 ELECTION RESULTS USING MULTIFACTOR TWITTER DATA ANALYSIS METHOD

İbrahim SABUNCU1, Mehmet Ali BALCI2, Ömer AKGÜLLER3

Abstract
In studies on election result prediction based on Twitter data, estimates were made using one of the factors such as the number of positive, negative, and neutral tweets posted about parties, the effect size of these tweets (the number of re-tweets), or the number of people who posted these tweets. However, no study was found that used all of these factors together. The goal of this study is to develop a new approach that takes into account all of the factors described and contributes to the literature in this context. A new multifactor model for the election result prediction based on Twitter data has been developed for this purpose. The model was tested by attempting to predict the results of the US 2020 elections in November, which had not yet taken place when the first version of this article was written. Also, a comparison has been made with alternative estimation approaches in the literature. Analyzes were made for approximately 10 million tweets collected between September 1 and October 21, 2020. As a result of the analysis, consistent with the results of the polls, Biden wins the election with a difference of 9.22% according to our method of estimating votes based on positive and negative tweet numbers, which are the current approaches in the literature. However, by using our multifactor model, the parameters for 3 November were calculated as -0.213423 for Democrats and 0.0455818 for Republicans. Based on these scores, it is concluded that the Republicans will win the election with a very small margin.

Introduction

Analysis of the relationship between social media data and electoral outcomes has started with the broad use of social media. Numerous studies have shown that social media has an impact on election outcomes in various countries (Bruns & Stieglitz, 2013) and can be used effectively to forecast election results (Awais et al., 2019; Bermingham & Smeaton, 2011; Bilal et al., 2019; Bose et al., 2019; Brito et al., 2019; Budiharto & Meiliana, 2018; Ceron et al., 2014; Elghazaly et al., 2016; Heredia et al., 2018; Jaidka et al., 2019; Jose & Chooralil, 2016; Khatua et al., 2015; Livne et al., 2010; Sang & Bos, 2012; Sharma & Moh, 2016; Singh et al., 2017; Skoric et al., 2012; Tsakalidis et al., 2015; Tumasjan et al., 2010; Xie et al., 2018; You et al., 2015).

Twitter is the most popular social media platform used for the prediction of election result studies using social media data. (Chauhan et al., 2020). Twitter is a social media application in many countries where millions of people share their views on different topics, politicians send political messages, and interact with the public (Grover et al., 2018). Also, voters often use Twitter to share their opinions on

---

1 Yalova University, Department of Industrial Engineering
2 Muğla Sıtkı Koçman University, Department of Mathematics
3 Muğla Sıtkı Koçman University, Department of Mathematics
candidates, politicians, and political debate. It has been reported that a substantial percentage of voters (for example, 25% of people in America (Grover et al., 2018)) are Twitter users, and an average of 34% of Twitter users share political opinions (Abanoz, 2020; Pew Research Center, 2012). The fact that the prediction of election results based on Twitter data is quicker and cheaper than the polls, means that Twitter can be used to monitor elections in real-time (Bansal & Srivastava, 2018). Another advantage is that the sample size provided by Twitter (the number of people whose political views are analyzed) is much higher than the surveys (Grover et al., 2018). These features have made Twitter an important source of data for election results prediction. Another benefit is that the sample size provided by Twitter (the number of people whose political views are analyzed) is much larger than that offered by the polls (Grover et al., 2018). These features have made Twitter an important data source for forecasting political outcomes.

A large number of studies have been conducted to forecast election outcomes via Twitter data on political patterns and elections in several different countries, such as Germany (Tumasjan et al., 2010), Ireland (Bermingham & Smeaton, 2011), USA (Choy et al., 2012), (Ramteke et al., 2016), (Wicaksono et al., 2017), (Anuta et al., 2017), (Grover et al., 2018), Dutch (Sang & Bos, 2012), (Sanders & Bosch, 2020), Pakistan (Mahmood et al., 2013), Colombia (Cerón-Guzmánn, 2016), UK (Burnap et al., 2016), Venezuelan (Castro et al., 2017), France (L. Wang & Gan, 2017), Indonesia (Ibrahim et al., 2016), (Budiharto & Meiliana, 2018), India (Bansal & Srivastava, 2018), (Singh et al., 2020), Poland (Jankowski, 2020), Spain (Grimaldi et al., 2020), Turkey (Abanoz, 2020).

Besides, studies have been conducted to acknowledge topics that are influential in election choices (Grover et al., 2018), forecast the political attitudes of Twitter users based on their (Makazhanov et al., 2014), (Toker et al., 2017). Some studies aim to establish the demographics of candidates' supporters by analyzing the images of supporters of election candidates on Twitter with image processing (Y. Wang et al., 2016).

In studies on election result prediction based on Twitter data, estimates were made using one of the factors such as the number of positive, negative, and neutral tweets posted about parties, the effect size of these tweets (the number of re-tweets), or the number of people who posted these tweets. However, no study was found that used all of these factors together. The goal of this study is to develop a new approach that takes into account all of the factors described and contributes to the literature in this context. A new model for the election result prediction based on Twitter data has been developed for this purpose. The model was tested by attempting to predict the results of the US 2020 elections in November, which had not yet taken place when the first version of this article was written. Also, a comparison has been made with alternative estimation approaches in the literature.

**Related Work**

Similar studies have been surveyed to develop the methodology to be used to predict election results. Data collection (tool/software, keywords, quantity), preliminary analysis (cleaning data), sentiment analysis, and election result prediction methods/tools/software used by these studies were briefly mentioned and implications were made from this information.

The study by Tumasjan et al. (Tumasjan et al., 2010) is one of the first studies on social media data and election results prediction. It was investigated whether Twitter was used as a forum for political negotiation and whether online messages on Twitter reflect offline political sentiment. To this end, they collected tweets containing the names of the 6 parties represented in the German parliament or the names of leading politicians to predict the German federal election results. Tweets were
automatically translated from German to English and Emotion analysis was performed with LIWC text analysis software (Yu et al., 2008). As a result, it has been determined that the ranking is the same according to the number of tweets mentioning the parties and the voting share in the election results. It has been stated that Twitter is indeed widely used for political conversations, and Twitter messages reasonably reflect the offline political landscape.

Bermingham and Smeaton (Bermingham & Smeaton, 2011) started to make predictions by analyzing Twitter social media data before the 2011 Irish General Election and started publishing the results with the tag “# GE11 Twitter Tracker”. They used both volume-based calculation and sentiment analysis in their studies where they assumed that the percentage of votes a party received was related to the volume of a relevant content on social media. Before the election, they collected 32,578 tweets about the election between 8 February and 25 February using party names and abbreviations related to the five main parties, as well as the election hashtag. The tweet was analyzed using the supervised learning sentiment analysis framework they developed. In this analysis, they used tokenizers developed by Laboreiro et al. (Laboreiro et al., 2010) optimized for user-generated content that preserves sociolinguistic features such as emoticons (“:”) and unusual punctuation marks (“!!!!”). They then estimated votes in three different ways: the ratio of the total number of tweets about the party to all tweets; The ratio of the number of positive tweets to all positive tweets is the ratio of the number of negative tweets to all negatives. Unexpectedly, the total tweet rate was more accurate than the positive tweet rate. They stated that they correctly predicted the number of votes for 5 parties with an average absolute error rate of 1.61% only by proportioning the number of tweets related to the party to the total number of tweets.

Sang and Bos (Sang & Bos, 2012) collected 64,395 tweets a week before the election to estimate the 2011 Dutch election results. They filtered tweets by applying a two-step process called normalization. They eliminated the tweets containing more than one party name, and they took the first tweet of each user, that is, one tweet for analysis. After these eliminations, 28,704 tweets remain. For analysis, they grouped the tweets just based on the party name in the tweets. The authors did the sentiment analysis manually. Then, they calculated the weight of emotions by comparing the number of non-negative tweets about a party to the total number of tweets about that party. They tried to estimate the number of seats and the rate of votes won by the parties by multiplying their sentiment analysis weight by the number of tweets posted about the party before the normalization process. They claim that the parties estimated their voting rate with an absolute difference of 17.4% in total.

In the first three studies examined (Tumasjan et al., 2010), (Bermingham & Smeaton, 2011), (Sang & Bos, 2012), it was seen that the election results were estimated by considering the number of tweets or positive tweets about the parties participating in the elections. However, in the study made by Choy et al. (Choy et al., 2012), the number of people who posted these tweets about the election was taken into account, not the number of tweets. They (Choy et al., 2012) collected 7,541,470 tweets over 3 months for the prediction of the 2012 USA Presidential Election result. They used AFINN (Nielsen, 2011) for sentiment analysis of the tweets they collected. After the sentiment analysis, they analyzed with the assumption that the person who positively tweeted about one of the presidential candidates supported that candidate. They multiplied the proportion of people who positively tweeted about a candidate by the proportion of internet users (using Twitter + not using Twitter) in the states. They also multiplied the proportion of support given to candidates in the states in the previous election, with the proportion of non-Internet users. They have calculated the total support for a candidate by adding both multiplication results. As a result of the analysis, they were able to correctly predict that Obama will win the 2012 election before the election.
Mahmood et al. (Mahmood et al., 2013), also took into account the number of tweeters to estimate the Pakistan 2013 election results. They identified 24 users who tweeted about the election and collected a total of 9000 tweets during the election period. They manually specified 40 words for or against the parties. According to these keywords in a tweet, they used RapidMiner grouping models to predict which party the tweet supports and which party it is against. They tried to estimate the vote rates of the parties by collecting the numbers of tweets for and against. However, they improperly predicted the election results with this small data set they gathered from very few users.

Another study that takes into account the Twitter users who tweeted about the election was made by Makazhanov et al. (Makazhanov et al., 2014). However, in this study, they developed a model not to predict the election results, but to predict the party that the Twitter user will vote for based on their Twitter posts. They collected tweets during the 10 days preceding the election, using 27 manually selected keywords such as party hashtags, party names, leader names, and general election hashtags related to the 2012 Albertan general election in Canada. As a result, they collected 181,972 tweets from 28,087 accounts. They removed the party candidates’ accounts and those with a communication language other than English from these accounts and identified 24,507 accounts. Distant supervision approach was used for the analyzes. Tweets by 252 candidates of the four major parties that entered the election were used for the training of the model. They developed a model that classifies Twitter users as a party’s supporters based on their characteristics such as retweeting, following, and interaction with the party. They also tested the model they developed for the 2013 Pakistani general elections. As a result, they stated that the model they developed could be used to predict the political preferences of Twitter users.

Ramteke et al. (Ramteke et al., 2016) again tried to estimate the election result based on the number of tweets. They collected tweets for Donald Trump and Hillary Clinton, two candidates who participated in the 2016 USA Election, on March 16-17, 2016, containing the names of the parties and candidates. They used 60 thousand tweets they obtained for analysis. They analyzed the tweets containing the hashtags used by the supporters of a party/candidate with the assumption that it could be positively tagged for that candidate/party. With this assumption, hashtags used at high frequency in the Tweet dataset (minimum 20 in this study) were manually tagged as positive or negative. For example, the #MakeAmericaGreatAgain hash has been labeled as positive support for Trump. In their analysis with the Phyton software, they first labeled the emotional polarity of the sentences with the VADER algorithm (Gilbert & Hutto, 2014). Then, by applying the Multinomial Naive Bayes and Support Vector machines machine learning models with the Python Scikit-learn package, they classified the tweets as positive and negative for the candidates. The estimation of the election result by dividing the total number of positive tweets sent about the candidate by the total number of tweets related to that candidate turned out to be incorrect.

In the thesis study by Guzmánn (Cerón-Guzmánn, 2016), tweet data were used to predict the 2014 Colombia election results. In the study, a machine learning approach was used to detect Spam tweet accounts. They emphasized that it is necessary to detect and weed out spam accounts for election result prediction analysis based on Twitter data, and to take into account the texts produced by users when performing sentiment analysis for tweets. As a result of the study, they stated that social media analysis has the potential to predict the voting rate of parties and to rank the highest-rated candidates accurately.

Burnap et al. (Burnap et al., 2016) collected tweets containing the names of the parties or candidates for more than 3 months to estimate the 2015 UK General Election results. They applied sentiment analysis to around 13 million tweets they obtained using the software developed by (Thelwall et al., 2010). They removed the tweets with content related to more than one party/candidate from the data.
They tried to estimate the vote ratio by collecting and rating all positive tweets about the party/leader. As a result of the study, although the parties incorrectly predicted the number of seats in the parliament, they correctly determined the voting ratio order.

Wicaksono et al. (Wicaksono et al., 2017) tried to predict the election results of states by collecting English tweets containing any keyword (e.g. republican, democrat, Hillary Clinton, Donald Trump) and location information about parties or candidates participating in the 2016 US presidential election. Using Sentiment140 tweet corpus (Go et al., 2009) that contain 1,600,000 data sets (800,000 training data for positive and negative emotion) and 497 test data (181 positives, 177 negatives, and 139 neutral) and abbreviations dictionary (Allacronyms, 2020), they performed sentiment analysis of tweets by 3 types of sentiment classifiers (Binarized Multinomial Naïve Bayes Classifier, SentiWordNet, and AFINN-111). As a result of the analysis, they considered the tweet about a party containing positive emotion as a positive vote for that party, and the one containing negative sentiment as a positive vote for the opposite party. This approach probably led to the erroneous calculation that the same person would vote positive votes for different parties due to different tweets. As a result of the study, it was wrongly predicted that Hillary Clinton (Democrat) won the election with a total of 253 electoral votes, as Donald Trump (Republican) received only 219 electoral votes.

Wicaksono (Wicaksono et al., 2017) and Ramteke (Ramteke et al., 2016) shows that it may not be sufficient to use only the number of total or positive tweets about parties to predict the election outcome, on the contrary of the first studies (Tumasjan et al., 2010), (Bermingham & Smeaton, 2011), (Sang & Bos, 2012). As a matter of fact, with the advancement of techniques such as artificial intelligence and machine learning, more advanced methods have been used for both sentiment analysis and voting rate estimation in election results prediction studies based on social media data. Another example of this kind of study is the work of Castro et al. (Castro et al., 2017) to predict the results of the 2015 Venezuelan Parliament election. In this study, they analyzed 60 thousand tweets published within the geographical borders of Venezuela one week before the election. In the study where they used social network analysis and unsupervised machine learning techniques, they were grouped according to states by using latitude and longitude information in tweets. They made predictions according to the frequency of the words they determined about the election in the tweets. Although they could only estimate the election results at a rate of 79.17%, they stated that they guessed the winning party correctly for 21 of the 24 states (87.50%) in Venezuela.

Anuta et al. (Anuta et al., 2017) conducted a study comparing the election forecast model based on Twitter data with polls. They used previously collected tweet data about the 2016 US elections in the study. They filtered the tweets in the data set according to the location (USA) and the language of the Twitter user. They saved 3 million tweets of 750,000 users to the PostgreSQL database for analysis. VADER (Gilbert & Hutto, 2014) application was used for sentiment analysis. According to the tweets of the people about the candidates, they decided whom they would vote for and estimated the votes by proportioning the number of people. As a result of the study, they stated that Twitter was more biased than surveys and the prediction was more erroneous with data based on tweets. However, this may be due to errors in their prediction methods based on Twitter data they use. As a matter of fact, if two candidates are mentioned in a tweet, one candidate may be praised and there may be a negative judgment against the other candidate. It is wrong to evaluate this tweet positively and analyze it as positive vote for both candidates.

In the study conducted by Wan and Gan (L. Wang & Gan, 2017), unlike previous studies, they estimated the election result taking into account not only the positive or negative but also the number of neutral tweets about the parties. First of all, they collected tweets about the 2017 French elections that were posted before the election and included the names of the candidates. They identified words in the
tweets that indicate a positive or negative emotion for a particular candidate. For example, they stated that the word Obama was positive for Macron, one of the candidates, because Obama published a video supporting Macron. They categorized these words for and against candidates. They categorized the tweets as positive or negative for candidates based on the emotion-expressing words in their content. Tweets that do not express emotion are categorized as neutral. They used the number of tweets to predict the election result. They proportioned the total number of positive tweets about the candidate to the number of positive and negative tweets about the candidate and multiplied by the ratio of all tweets (including neutral) about the candidate among all tweets in the sample. They estimated the daily election results according to the tweets they collected using the formulas they specified. On the last day, they predicted the election results with only a 2% margin of error. They stated that their vote rate estimates, taking into account the neutral tweets, were more successful than the methods considered only for positive and negative tweets.

Toker et al. (Toker et al., 2017), similar to the study of Makazhanov et al. (Makazhanov et al., 2014), tried to determine the political tendency of Twitter users. They stated that, according to the words used in the tweets of Twitter users, they determined which party supporters (political tendencies) they were. For their analysis, they selected 1011 Twitter users by random sampling from among the followers of a comedian who used their real name and real picture and who posted at least 130 tweets in the last three months. Depending on their tweets or retweets that contain words related to Turkey 2015 June elections, these users' political tendencies were tried to be identified. The Twitter messages of 364 people selected among these users were re-analyzed for the 2015 November elections and investigated the changes in the political tendency. They stated that the political tendency of 68.2% of the sample they determined was determined.

Grover et al. (Grover et al., 2018) collected a total of 784,153 tweets about the election in three months in their study after the 2016 USA election. Tweets were collected using keywords related to candidates and parties. Besides, all the tweets of the two candidates of the election (Hillary Clinton "and" Donald Trump ") were collected. Tweet collecting process and sentiment analysis were done in the R programming language with syuzhet, lubridate, and dplyr libraries. For topic modeling, tm and topicmodels libraries of R were used. According to this study, discussions on Twitter can polarize users, that is, they affect them. They analyzed the polarization using the Newman model (Newman & Sheth, 1985). As a result of the geographic analysis of the tweets, they also found that Twitter causes ideas and ideologies to affect other users (acculturation). Acculturation has been defined as the change in preferences within an individual when interacting with individuals or groups with a different cultural background (Redfield et al., 1936). Another interesting result of the study is that although the net positive (positive-negative) tweets about Clinton are higher than Trump's, the average number of likes for Trump's tweets is almost 3.8 times that of Clinton. Now that Trump won the election, the conclusion to be drawn from this study is that it would be useful to include not only the amount of tweets, but also the number of likes and retweets in the analysis.

Another study that takes into account the number of people who tweeted about the election is the study of Bansal & Srivastava (Bansal & Srivastava, 2018). Using the R programming language, they had collected more than 300,000 tweets related to India Uttar Pradesh legislative elections. They used words including party names, party leader names, election campaign slogans, and hashtags as keywords for collecting related tweets. They performed sentiment analysis on these tweets with a method they developed themselves and called Hybrid Topic-Based Sentiment Analysis (HTBSA). With this method, a sentiment score was calculated for each tweet. They estimated the election result by taking into account the total number of Twitter users who had sent positive tweets about the parties. The fundamental assumption in this calculation is that a person’s positive tweet about a party is the
intention to vote for that party. As a result of this study that had conducted after the election, they reported that they predicted the election results with an average absolute error of 8.4%. They also claimed that the HTBSA method they developed was more successful than other lexicon-based sentiment analysis methods.

Even after 8 years have passed since the first studies (Tumasjan et al., 2010), there are still studies that take into account only positive tweet counts. For example, in the study conducted by Budiharto & Meiliana (Budiharto & Meiliana, 2018), using the R programming language, tweets containing the determined hashtag (#) related to the Indonesian 2018 selection were collected. Sentiment analysis was applied to the pre-processed tweets using the TextBlob library. They predicted with a simple approach that the candidate with a high number of positive tweets will win the election. They stated that their prediction turned out to be correct as a result of their choice.

In the social media application (Singh et al., 2020), which was made to predict the Indian Punjab state election results, data was taken from Twitter using hashtags related to the parties. Only 9157 tweets were collected and all tweets were translated into English. Before the analysis, punctuation marks, numbers, web page links, extra spaces were removed from the content of the tweets as a pre-process. Tweets are associated with parties by hashtag (#) or account name (@) found in tweets. Network analysis was performed to visualize how many tweets were posted about which party. They performed sentiment analysis on the tweets. Then they subtracted the total number of negative tweets from the total number of positive tweets about a party and proportioned this figure by subtracting the total number of negative tweets from the total number of positive tweets of all parties. They accepted the rate they found as the voting rate. From this ratio, they used simple linear regression to estimate the number of seats in the council. They state that the estimation of the number of seats they obtained is correct as a result of their analysis using the voting rates and the number of seats of previous years.

In the study by Grimaldi et al. (Grimaldi et al., 2020) sentiment analysis and related tweet amounts were used to predict the 2019 Spain election results. One million tweets covering parties, candidates, and election-related hashtags have been collected from Twitter. Before analysis, they deleted unwanted phrases, words, URLs, numbers, and dates from these tweets. Using the frequencies of the expressions in the tweets, they extracted the attributes with the R software. They used five different machine learning algorithms in their analysis and compared their performance. They obtained the most accurate selection result estimation with the kNN (k-Nearest Neighbor) classification algorithm.

In the study of Sanders and Bosch (Sanders & Bosch, 2020), it was investigated which type of tweets (making a positive comment, recommending a candidate, etc.) are more effective in predicting the election result. For this purpose, 17,000 tweets were manually tagged by 500 people at least three times for 10 different categories. In other words, each tweet is tagged for 10 different categories such as mockery, recommendation, positive or negative according to the word it contains. Then, using these 10 categories (annotations), the election results were tried to be predicted. It was analyzed which of these 10 categories correctly predicted the election results with the lowest margin of error. This study was conducted for three different elections in the Netherlands. As a result, a single category with the lowest error could not be determined among all three selections. Therefore, the most useful category of tweet suggestions (eg positive, suggestive, etc.) could not be submitted for the prediction of election results.

In a study conducted for the elections in Turkey (Abanoz, 2020), changes in the interest of Twitter users to the parties have been estimated depending on the number of daily tweets. During, the general election in 2015, the referendum in 2017, and the general election in 2018, tweets containing the names of the parties or election slogans or the names of their leaders were collected. More than 3
million tweets were collected during each election period. In the analysis, they only calculated the daily changes in the number of tweets. But they did not apply sentiment analysis to tweets. They group the tweets as positive or negative according to the hashtag/keyword they contain. They calculated the interest of Twitter users by looking at the number of tweets posted. As a result of the study, they stated that they could detect the change of Twitter users’ interest in parties on daily basis.

In the study of Chauhan et al. (Chauhan et al., 2020), previous works made for the prediction of the election result using social media were examined. Data collection and analysis methods used in these studies were compiled. Suggestions have been made for researchers to conduct similar studies. Within the scope of the article, 38 articles related to the election result estimation based on social media data published between 2010 and 2019 were examined. It was seen that in 23 of these articles, data were collected from Twitter social media. It was mentioned that sentiment analysis and volumetric approach were used as the analysis method for the selection outcome prediction. It was stated that in 24 of the studies, the election results were correctly estimated. While most of the studies (29 of them) were conducted after the election, only a few (9 of them) were carried out before the election. It is emphasized that it is more meaningful to state the election results before the elections. It was reported that to obtain a more accurate election result prediction, attention should be paid to the geographical location and age limit to determine and eliminate non-voters’ tweets. It was also emphasized that it would be beneficial to remove spam or mocking messages and tweets belonging to bot accounts. It was pointed out that analyzing different social media data such as Facebook and Instagram in addition to Twitter will increase the prediction performance. As an analysis method, it was mentioned that deep learning gives more accurate results.

As a result of the literature review, the factors that are found to be effective in the election result prediction based on social media data are the number of positive, negative, and neutral tweets, the effect size of these tweets (the number of re-tweets), and the number of people who posted these tweets. It is thought that a new method that takes all these factors into account will give more effective results than existing methods and contribute to the literature in this context. Below, a new prediction model based on Twitter data has been developed for this purpose. The model has been tested by trying to predict the result of the USA 2020 November elections, which had not yet occurred at the time of publishing the first version of this article.

Methodology

According to the previous studies, in social media analytics applications, the number of people tweeting for/against a party; the number of positive, negative, and even neutral tweets about a party; and the amount of retweeting of these tweets are important factors for predicting election results. A prediction model that takes all these factors into account could not be found in the literature. So, in this study, a new election result prediction model based on social media data was created, which takes all the aforementioned factors into account.

Election results prediction studies are generally carried out after the election, but publishing them before the election is more meaningful (Chauhan et al., 2020). Therefore, to evaluate the success of it, the newly developed model was used to predict the result of the general election of the USA in November 2020 which has not yet been realized at the time of publishing the first version of this article.

The selection result estimation obtained with the newly developed model has been compared with other estimation models in the literature and survey results. The compared models are the prediction model based on the number of tweeter users (Bansal & Srivastava, 2018), (Anuta et al., 2017), (Choy et al., 2012); prediction model based on the number of positive and negative tweets (Singh et al.,
Data Collection

In the examined studies, it was observed that the keywords used in collecting the election-related tweets included the names of the parties, their abbreviations, the names of the party candidates, and the election slogans. So, in this study, data were collected using similar concepts. Tweets containing at least one of the 29 keywords about the four parties that received the most votes in the last election, were collected. These keywords are given in Table 1. Tweets have been collected since July 1, 2020. It is planned to continue the data collection process until November 14, 2020, ten days after the election.

The number of tweets collected is more than 10 million. The election results estimation study made using more than this amount of tweets in the literature could only be identified in 3 (Burnap et al., 2016; Ceron et al., 2014; Shi et al., 2012). Therefore, the number of tweets collected is considered to be sufficient for analysis.

In some of the reviewed studies, the election-related tweets were collected in the last month or less before the election (Bansal & Srivastava, 2018), (Castro et al., 2017), (Sang & Bos, 2012), (Bermingham & Smeaton, 2011), and even in some studies (Ramteke et al., 2016), it was seen that they gathered just a day before the election. In this study, data were collected four months before the election. This long time allows a better analysis of the changing popularity of parties.

The collected tweets are divided into four different groups as the first 40 days (from July 1 to August 12, 2020), the next 40 days (from August 13 to September 25), the last 40 days (from September 26 to November 3) and the first 10 days after the elections (from November 4 to 14). The analyzes were applied to these groups separately. By comparing the results of the analysis, the changes in the period before and after the election were also analyzed.

Sentiment Analysis

Sentiment analysis is an automatic text processing technique used to measure views and attitudes. Although widely used in social media data to assess general perceptions of products and brands, it can provide insight into events and issues. Rather than relying on manual techniques to identify positive or negative themes, sentiment analysis compares terms to a vocabulary dictionary with predetermined

| Category                    | Keyword(s)                                                                 |
|-----------------------------|-----------------------------------------------------------------------------|
| US November 2020 Election   | #USAelection OR #NovemberElection                                            |
| Democratic Party            | @DNC OR @TheDemocrats OR Biden OR @JoeBiden OR "Our best days still lie ahead" OR "No Malarkey!" |
| Green Party                 | @GreenPartyUS OR @TheGreenParty OR “Howie Hawkins” OR @HowieHawkins OR “Angela Walker” OR @AngelaNWalker |
| Libertarian Party           | @LPNational OR “Jo Jorgensen” OR @Jorgensen4POTUS OR “Spike Cohen” OR @RealSpikeCohen |
| Republican Party            | #MAGA2020 OR @GOP OR Trump OR @POTUS OR @realDonaldTrump OR Pence OR @Mike_Pence OR @VP OR "Keep America Great" |
emotion scores. In addition to the wide variety of methods utilizing different dictionaries and scales, the standard "Syuzhet" R package is used in this study.

The package comes with four sentiment dictionaries and provides a method for accessing the robust, but computationally expensive, sentiment extraction tool developed in the NLP group at Stanford (Jockers, 2017). The default “Syuzhet” lexicon was developed in the Nebraska Literary Lab under the direction of Matthew L. Jockers is used in this study. Within each tweet, each word is evaluated independently and a score is given, and a total score is returned for each tweet. Positive scores indicate positive attitudes, negative points indicate negative attitudes, and words not found in the corresponding dictionary get zero.

**Multifactor Twitter Data Analysis Method**

Our most important contribution in this study is to analyze by giving weight not only to the number of views of Twitter users but also to the views of the most influential users in the network structure formed by users speaking about US Elections 2020.

For the network creation process, we first determine the users who have interaction among them throughout the mentioning or direct messages. Such interactions are extracted from the data set and a weighted directed network is formed. Each user has a fixed User-Id; hence, we name each vertex after the User-Ids. The weight function is chosen to be the number of Retweets each interaction has.

The most influential users in such a network corresponding to the vertices throughout their PageRank index. PageRank centrality in a network is a measure that can be summarized as; a vertex is important if it is linked from other important and link parsimonious vertices if it is highly linked (Page et al, 1999). Three important factors determine the PageRank of a vertex. The first factor is the number of edges the vertex receives. This factor emerges from the idea of the more edges a vertex attracts, the more it is perceived. The second factor is the edge propensity of the vertices which depends on the idea of the number of edges coming from parsimonious vertices which are more valuable than those emanated by spendthrift ones.

Let $A = (a_{ij})$ be the adjacency matrix of a directed network. Then, the PageRank centrality $x_i$ of the user, $i$ can be defined by

$$x_i = \alpha \sum_k \frac{a_{ki}}{d_k} x_k + \beta,$$

where $\alpha$ is a damping factor, $\beta$ is an exogenous variable, $d_k$ is out-degree of vertex $k$. Then, in matrix form we have

$$x = \alpha x D^{-1} A + \beta,$$

where $\beta$ is a positive constant vector which is called the personalization vector, $D^{-1}$ is a diagonal matrix whose $i$-the diagonal element equals to $1/d_i$. For large networks like the Twitter network, we prefer to use the power method for the computation of PageRank (Maehara et al, 2014; ). Moreover, the damping factor $\alpha$ should be chosen between 0 and $1/\rho$, where $\rho$ is the largest eigenvalue of the matrix $D^{-1}A$, to guarantee the convergence of the measure (Kamwar et al, 2004).

In the network of users speaking about American elections, the number of retweets of tweets as well as measurements of users' activity is an indicator of participating in positive or negative opinions (Boyd et al, 2010). Considering the emotion score of a tweet posted by a user, the centrality of the user in
the interaction network, and the number of retweets of the tweet, a score as follows is determined for each user in this study:

\[ MFS_i = 1000x_i \sum S_jR_j, \]

where \( x_i \) is the PageRank centrality of user \( i \), \( S_j \) and \( R_j \) are the sentiment score and retweet number of the \( j \)-th tweet by user \( i \), respectively, and 1000 is a normalization coefficient. In the data set, the retweet numbers of direct interaction tweets are found to be mostly 0. Hence, we consider \( R_j = 1 \) for these cases.

Unlike the methods in the literature, the autoregressive fractionally integrated moving average model (FARIMA) is used in the study for the estimation of the daily average \( MFS_i \) scores on 3rd November 2020 in the datasets that we grouped as Democrats and Republicans between 1st September 2020 and 21st October 2020, instead of the \( MFS_i \) totals. Thus, it is expected to predict the election outcome depending on the multifactorial development of the discourse for both groups.

FARIMA models generalize ARIMA (autoregressive integrated moving average) models by allowing non-integer values of the differencing parameter \( d \). Formally, let \( \{X_t\} \) be a stationary process such that

\[ \phi(B)(1 - B)^dX_t = \theta(B)Z_t, \quad \{Z_t\} \sim WN(0, \sigma^2), \]

where

\[ (1 - B)^d = \sum_{k=0}^{\infty} \left( \frac{d}{k} \right) (-1)^k B^k \]

and \( |d| < 0.5 \). Hence, by considering \( \langle MFS_t \rangle \sim X_t \) at day \( t \) between 1st September 2020 and 21st October 2020, we can forecast the \( \langle MFS_t \rangle \) on 3rd November 2020 with convenient parameter estimations.

**Findings**

**Descriptive Statistical Analysis**

The datasets are grouped as Democrats and Republicans from September 1, 2020, to October 21, 2020. A total of 2840373 tweets were posted for Democrats and 6691801 for Republicans. Among these total tweets, the positive tweet rate for Democrats is 42.96%, the negative tweet rate is 42.78%, the neutral tweet rate is 14.26%, for Republicans the positive tweet rate is 39.58%, the negative tweet rate is 53.54%, and the neutral tweet rate is 6.88%. See Figure 1.

Users who speak English about the elections on Twitter have the 100 highest PageRank score in the interaction network have statics as follows: for Democrats, the rate of positive tweets is 42.4%, the rate of negative tweets 43%, the rate of neutral tweets 43%, for Republicans the rate of positive tweets 37.81%, and the rate of negative tweets 50.22%, neutral tweet rate 11.97%. Besides, it has been observed that active users who speak for Democrats and Republicans use strong rhetoric, especially when speaking positive or negative about Republicans. See Figure 2.
Figure 1: The numbers of positive, negative, and neutral tweets
Figure 2. Scored sentiments of the users with the first hundred highest PageRank centrality.
USA November 2020 Election Results Prediction Analysis

The $\langle MFS_t \rangle$ values for 1st September – 21st October are presented in Figures 3 and 4. The fitted FARIMA models have parameters (1,-0.103,0) for Democrats and (0,-0.0254,0) for Republicans. The forecasts for 3rd November read as -0.213423 for Democrats and 0.0455818 for Republicans.
Alternative Prediction Method using Positive and Negative Tweet Numbers

According to Wicaksono et al. (Wicaksono et al., 2017), a positive tweet about a candidate indicates the intention to vote for that candidate, and a negative tweet indicates the intention to vote for an opposition candidate. Using this approach, as an alternative, the ratio of the total number of positive tweets sent for a party and negative tweet sent against the opposing party to the total number of all positive and negative tweets is used in the vote ratio estimation. According to this calculation in the literature, Biden will win the election with a margin of 9.22%. So, if it is assumed that the total vote ratio of Republicans and Democrats will be the same as in the previous (2016) election, the voting rate
for Democrats was estimated at 51.76% and for Republicans at 42.54%. This result is consistent with the polls.

References

Abanoz, E. (2020). A Twitter-Based Analysis Of Hashtag And Mention Actions As An Indicator Of Turkish General Elections’ Outcomes. Akdeniz Üniversitesi İletişim Fakültesi Dergisi, 33, 73–90.

Allacronyms. (2020). Social Networking Acronyms and Abbreviations. https://www.allacronyms.com/social_networking/abbreviations

Anuta, D., Churchin, J., & Luo, J. (2017). Election bias: Comparing polls and Twitter in the 2016 us election. ArXiv Preprint ArXiv:1701.06232.

Awais, M., Hassan, S. U., & Ahmed, A. (2019). Leveraging big data for politics: predicting general election of Pakistan using a novel rigged model. Journal of Ambient Intelligence and Humanized Computing, 0123456789. https://doi.org/10.1007/s12652-019-01378-z

Bansal, B., & Srivastava, S. (2018). On predicting elections with hybrid topic based sentiment analysis of tweets. Procedia Computer Science, 135, 346–353. https://doi.org/10.1016/j.procs.2018.08.183

Bermingham, A., & Smeaton, A. F. (2011). On Using Twitter to Monitor Political Sentiment and Predict Election Results. Proceedings Of the Workshop on Sentiment Analysis Where AI Meets Psychology (SAAIP), 2–10.

Bilal, M., Asif, S., Yousuf, S., & Afzal, U. (2019). 2018 Pakistan General Election: Understanding the Predictive Power of Social Media. 12th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics, MACS 2018 - Proceedings, 1–6. https://doi.org/10.1109/MACS.2018.8628445

Bose, R., Dey, R. K., Roy, S., & Sarndar, D. (2019). Analyzing Political Sentiment Using Twitter Data BT - Information and Communication Technology for Intelligent Systems (S. C. Satapathy & A. Joshi (eds.); pp. 427–436). Springer Singapore.

Boyd, D., Golder, S., & Lotan, G. (2010, January). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In 2010 43rd Hawaii international conference on system sciences (pp. 1-10). IEEE.

Brito, K., Paula, N., Fernandes, M., & Meira, S. (2019). Social Media and Presidential Campaigns – Preliminary Results of the 2018 Brazilian Presidential Election. Proceedings of the 20th Annual International Conference on Digital Government Research, 332–341. https://doi.org/10.1145/3325112.3325252

Bruns, A., & Stieglitz, S. (2013). Towards more systematic Twitter analysis: metrics for tweeting actweetsies. International Journal of Social Research Methodology, 16(2), 91–108. https://doi.org/10.1080/13645579.2012.756095

Budiharto, W., & Meiliana, M. (2018). Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis. Journal of Big Data, 5(1), 51. https://doi.org/10.1186/s40537-018-0164-1

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 Studies, 41. https://doi.org/10.1016/j.electstud.2015.11.017
Castro, R., Kuffó, L., & Vaca, C. (2017). Back to #6D: Predicting Venezuelan states political election results through Twitter. *2017 4th International Conference on EDemocracy and EGovernment, ICEDEG 2017*. https://doi.org/10.1109/ICEDEG.2017.7962525

Cerón-Guzmánn, J. A. (2016). A Sentiment Analysis Model of Spanish Tweets. Case Study: Colombia 2014 Presidential Election. In *Universidad Nacional de Colombia* ....

Ceron, A., Curini, L., & Iacus, S. M. (2014). Using Sentiment Analysis to Monitor Electoral Campaigns: Method Matters—Evidence From the United States and Italy. *Social Science Computer Review, 33*(1), 3–20. https://doi.org/10.1177/0894439314521983

Chauhan, P., Sharma, N., & Sikka, G. (2020). The emergence of social media data and sentiment analysis in election prediction. *Journal of Ambient Intelligence and Humanized Computing, 1*–27. https://doi.org/10.1007/s12652-020-02423-y

Choy, M., Cheong, M. L. F., Ma, N. L., & Koo, P. S. (2012). US Presidential Election 2012 Prediction using Census Corrected Twitter Model. *Research Collection School Of Information Systems, 1*–12. http://arxiv.org/abs/1211.0938

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1*(Mlm), 4171–4186.

Elghazaly, T., Mahmoud, A., & Hefny, H. A. (2016). Political Sentiment Analysis Using Twitter Data. *Proceedings of the International Conference on Internet of Things and Cloud Computing*. https://doi.org/10.1145/2896387.2896396

Gilbert, C. H. E., & Hutto, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) Http://Comp. Social. Gatech. Edu/Papers/icwsm14. Vader. Hutto. Pdf, 81, 82.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford, 1*(12), 2009.

Grimaldi, D., Diaz, J., & Arboleda, H. (2020). Inferring the votes in a new political landscape. The case of the 2019 Spanish Presidential elections. *Preprint from Research Square*. https://doi.org/10.21203/rs.3.rs-16463/v2

Grover, P., Kar, A. K., Dwivedi, Y. K., & Janssen, M. (2018). Polarization and acculturation in US Election 2016 outcomes – Can Twitter analytics predict changes in voting preferences. *Technological Forecasting and Social Change, September*, 1–23. https://doi.org/10.1016/j.techfore.2018.09.009

Heredia, B., Prusa, J. D., & Khoshgoftaar, T. M. (2018). Social media for polling and predicting United States election outcome. *Social Network Analysis and Mining, 8*(1), 1–16. https://doi.org/10.1007/s13278-018-0525-y

Ibrahim, M., Abdillah, O., Wicaksono, A. F., & Adriani, M. (2016). Buzzer Detection and Sentiment Analysis for Predicting Presidential Election Results in a Twitter Nation. *Proceedings - 15th IEEE International Conference on Data Mining Workshop, ICDMW 2015*, 1348–1353. https://doi.org/10.1109/ICDMW.2015.113

Jaidka, K., Ahmed, S., Skoric, M., & Hilbert, M. (2019). Predicting elections from social media: a three-country, three-method comparative study. *Asian Journal of Communication, 29*(3), 252–273. https://doi.org/10.1080/01292986.2018.1453849
Jankowski, R. (2020). *Predicting election polls results using machine learning tools* [Pracownia Fizyki w Ekonomii i Naukach Społecznych]. http://repo.bg.pw.edu.pl/index.php/en/r#/info/bachelor/WUT4a113c9f11ad427b98d0d9aa76c987be/

Jose, R., & Chooralil, V. S. (2016). Prediction of election result by enhanced sentiment analysis on Twitter data using classifier ensemble Approach. *Proceedings of 2016 International Conference on Data Mining and Advanced Computing, SAPIENCE 2016, November*, 64–67. https://doi.org/10.1109/SAPIENCE.2016.7684133

Jockers, M. (2017) “Syuzhet Package.” Available from: https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html. [Accessed 28th October 2020].

Kamvar, S., Haveliwala, T., & Golub, G. (2004). Adaptive methods for the computation of PageRank. *Linear Algebra and its Applications*, 386, 51–65.

Khatua, A., Khatua, A., Ghosh, K., & Chaki, N. (2015). Can #Twitter-Trends predict election results? Evidence from 2014 Indian general election. *Proceedings of the Annual Hawaii International Conference on System Sciences, 2015-March*, 1676–1685. https://doi.org/10.1109/HICSS.2015.202

Laboreiro, G., Sarmento, L., Teixeira, J., & Oliveira, E. (2010). Tokenizing micro-blogging messages using a text classification approach. *Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data*, 81–88.

Leamon, E. J., & Bucelato, J. (2017). Federal Elections 2016. In *Federal Election Commission*. https://doi.org/10.2307/j.ctvk8w129.20

Livne, A., Simmons, M. P., Adar, E., & Adamic, L. a. (2010). The Party is Over Here : Structure and Content in the 2010 Election. *October*, 161(3), 201–208. https://doi.org/10.1007/s00024-003-2459-0

Maehara, T., Akiba, T., Iwata, Y., & Kawarabayashi, K. I. (2014). Computing personalized PageRank quickly by exploiting graph structures. *Proceedings of the VLDB Endowment*, 7(12), 1023-1034.

Mahmood, T., Iqbal, T., Amin, F., Lohanna, W., & Mustafa, A. (2013). Mining Twitter big data to predict 2013 Pakistan election winner. *International Multi Topic Conference (INMIC)*, 49–54.

Makazhanov, A., Rafiei, D., & Waqar, M. (2014). Predicting political preference of Twitter users. *Social Network Analysis and Mining*, 4(1). https://doi.org/10.1007/s13278-014-0193-5

Newman, B. I., & Sheth, J. N. (1985). A model of primary voter behavior. *Journal of Consumer Research*, 12(2), 178–187.

Nielsen, F. Å. (2011). *AFINN sentiment analysis in Python: Wordlist-based approach for sentiment analysis*. Technical University of Denmark. https://github.com/fnielsen/afinn

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web*. Stanford InfoLab.

Pew Research Center. (2012). *Social Networking Popular Across Globe*. https://www.pewresearch.org/global/2012/12/12/social-networking-popular-across-globe/

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 9.

Ramteke, J., Shah, S., Godhia, D., & Shaikh, A. (2016). Election result prediction using Twitter sentiment analysis. *2016 International Conference on Inventive Computation Technologies*
Redfield, R., Linton, R., & Herskovits, M. J. (1936). Memorandum for the study of acculturation. *American Anthropologist, 38*(1), 149–152.

Sanders, E., & Bosch, A. van den. (2020). Optimising Twitter-based Political Election Prediction with Relevance and Sentiment Filters. *Proceedings of The 12th Language Resources and Evaluation Conference*, 6158–6165.

Sang, E. T. K., & Bos, J. (2012). Predicting the 2011 dutch senate election results with Twitter. *Proceedings of The Workshop on Semantic Analysis in Social Media*, 53–60.

Skoric, M., Poor, N., Achananuparp, P., Lim, E.-P., & Jiang, J. (2012). Tweets and Votes: A Study of the 2011 Singapore General Election. *2012 45th Hawaii International Conference on System Sciences*, 2583–2591. https://doi.org/10.1109/HICSS.2012.607

Tokesan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. *Fourth International AAAI Conference on Weblogs and Social Media*, 37(2), 455–479. https://doi.org/10.15581/009.37.2.455-479

Wang, L., & Gan, J. Q. (2017). Prediction of the 2017 French election based on Twitter data analysis. *2017 9th Computer Science and Electronic Engineering (CxEEC)*, 89–93.

Wang, Y., Li, Y., & Luo, J. (2016). Deciphering the 2016 U.S. presidential campaign in the Twitter sphere: A comparison of the Trumpists and Clintonists. *Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016*, 723–726.
Xie, Z., Liu, G., Wu, J., & Tan, Y. (2018). Big data would not lie: prediction of the 2016 Taiwan election via online heterogeneous information. *EPJ Data Science*, 7(1). https://doi.org/10.1140/epjds/s13688-018-0163-7

You, Q., Cao, L., Cong, Y., Zhang, X., & Luo, J. (2015). A Multifaceted Approach to Social Multimedia-Based Prediction of Elections. *IEEE Transactions on Multimedia, 17*(12), 2271–2280. https://doi.org/10.1109/TMM.2015.2487863

Yu, B., Kaufmann, S., & Diermeier, D. (2008). *Exploring the characteristics of opinion expressions for political opinion classification.*