Computational and Mathematical Organization Theory
https://doi.org/10.1007/s10588-023-09376-5

ORIGINAL PAPER

280 characters to the White House: predicting 2020 U.S. presidential elections from twitter data

Rodrigue Rizk[1] · Dominick Rizk[2] · Frederic Rizk[2] · Sonya Hsu[2]

Accepted: 6 March 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

This nation-shaping election of 2020 plays a vital role in shaping the future of the U.S. and the entire world. With the growing importance of social media, the public uses them to express their thoughts and communicate with others. Social media have been used for political campaigns and election activities, especially Twitter. The researchers intend to predict presidential election results by analyzing the public stance toward the candidates using Twitter data. Previous researchers have not succeeded in finding a model that simulates well the U.S. presidential election system. This manuscript proposes an efficient model that predicts the 2020 U.S. presidential election from geo-located tweets by leveraging the sentiment analysis potential, multinomial naive Bayes classifier, and machine learning. An extensive study is performed for all 50 states to predict the 2020 U.S. presidential election results led by the state-based public stance for electoral votes. The general public stance is also predicted for popular votes. The true public stance is preserved by eliminating all outliers and removing suspicious tweets generated by bots and agents recruited for manipulating the election. The pre-election and post-election public stances are also studied with their time and space variations. The influencers' effect on the public stance was discussed. Network analysis and community detection techniques were performed to detect any hidden patterns. An algorithm-defined stance meter decision rule was introduced to predict Joe Biden as the President-elect. The model's effectiveness in predicting the election results for each state was validated by the comparison of the predicted results with the actual election results. With a percentage of 89.9%, the proposed model showed that Joe Biden dominated the electoral college and became the winner of the U.S. presidential election in 2020.

Keywords Social cybersecurity · Network analysis · Social media analytics · Sentiment analysis · Elections · Misinformation · Fake news

✉ Rodrigue Rizk
rodrigue.rizk@usd.edu

Extended author information available on the last page of the article

Published online: 28 March 2023

1 Introduction

2020 is considered the year of challenges and crises that have been striking the whole world, starting with the coronavirus outbreak, riots and protests, wildfires, and ending with the U.S. presidential election. This nation-shaping election plays an important role in shaping the future of the United States of America and the entire world. With the growing importance of social media in human life, people use it as a social platform for expressing their thoughts and communicating with each other. However, these platforms have started to be used by politicians for political purposes, especially Twitter, which has surpassed 340 million users (Sehl 2021) with 500 million tweets per day (Sayce 2021). Candidates find it as a medium where they can promote their election campaign. Thus, knowing people's stances toward candidates affects the election results since candidates will know the public stance toward them and in what ways they can improve their reputation before the election day. The race of election campaigns on Twitter leads researchers to predict election results through Twitter data by analyzing the public stance toward the candidates. (Shi et al. 2012; Tumasjam et al. 2010).

Researchers have been applying sentiment analysis on tweets to know people's stances toward the candidates to predict the election results (Yaqub et al. 2020; Liu et al. 2021; Abroms and Lefebvre 2009; Nausheen and Begum 2018; Oikonomou and Tjortjis 2018; Rathi et al. 2018). However, none has succeeded in finding a model that resembles the U.S. presidential election system since it is a bipartite system and based on the electoral college. Two candidates represent the bi-parties and whoever wins most states wins the election. In other words, knowing the public stance toward a candidate will not be beneficial for knowing who will get most of the electoral votes and be the election's winner. Still, it will help in knowing who will get most of the popular votes, which does not ensure winning the election. Since the U.S. election is a state-based system, it is appropriate to find a model that finds public stance through analyzing tweets and takes into consideration the geolocation of Twitter's users, focusing more importantly on the state location. By doing so, the public stance for each state, the winner of electoral votes, and the winner of the election are determined. Studying the pre-election and post-election stances is very important to know how the public's stance changes with the newly elected President.

Many factors affect the public stance, one of which is the influencers and their effects on the public stance. Since the influencers have a large popular base, they are considered key elements in any election for supporting a candidate, promoting him, and even helping him in his election campaign. They play a key role in biasing the public stance toward a specific candidate, thus affecting the election result. One could take as an example Lady Gaga and her support for Joe Biden. Therefore, it's interesting to study the effect of influencers on public stance.

Not all people's opinions on Twitter are innocent; some tweets might be generated by bots; others could be generated by an agent aimed at manipulating the public stance or even spreading fake news or hatred. So, it's crucial to eliminate these suspicious tweets in order to know the true public stance.

Our contributions are summarized as follows. The general public stance for popular votes is predicted along with the state-based public stance for electoral votes. The pre-election and the post-election public stances and their variation with space and time are determined. The influencers' effect on the public stance is also considered. Moreover, fraud detection, bot activity, and election manipulation are investigated. Furthermore, network analysis and community detection are performed. In addition, a stance meter decision rule algorithm is also introduced. Finally, the sentiments of the tweets are validated by the on-ground public opinion. To highlight further our contributions and the novelty of our work, a comparative analysis with the most closely related works (Heredia et al. 2018; Yaqub et al. 2020, Liu et al. 2021) in the literature, that covers the U.S. presidential elections, is performed. Existing efforts (Heredia et al. 2018; Avello et al. 2011; Metaxas et al. 2011) have been unable to recreate the electoral college system, which is the core of the U.S. presidential elections. On the other hand, our work employs a hybrid classification method for sentiment analysis. It is more suited to the state-based electoral college system, particularly in forecasting the results of electoral votes. It takes into account both spatial and temporal differences from pre-election to post-election periods for all states in the United States. As a consequence, it can be used in future electoral campaigns as a supplement to projecting the results of the U.S. presidential elections.

To address the limitations of existing works (Heredia et al. 2018; Avello et al. 2011; Metaxas et al. 2011), an efficient framework for predicting the 2020 U.S. presidential election from tweets is proposed. It takes into consideration the geolocation of the users while leveraging the sentiment analysis potential, multinomial naive Bayes classifier, and machine learning. The model also predicts the winner of the popular votes and the winner of the electoral votes. All outliers are eliminated by removing suspicious tweets to preserve the true public stance. A study of the spatiotemporal change of the public stance before and after the election for each state is performed. Moreover, the effect of influencers on the public stance is also investigated. People's sentiments on Twitter are compared with their on-ground opinions by validating the outcomes with the election results. A stance meter decision rule algorithm is introduced to project the winner of the presidential election. Network analysis and community detection techniques are performed to detect any hidden patterns.

The main motivation behind using sentiment analysis for predicting elections is that sentiment analysis has been proven to be a viable technique for predicting people's sentiments with high accuracy and the closest estimate of the on-ground public stance (Nausheen and Begum 2018; Ramzan et al. 2017; Ayata et al. 2017). Moreover, the intuition behind adopting machine learning techniques, especially the multinomial naïve Bayes, is that they have higher success rates than other traditional algorithms (Zhang 2004) in sentiment classification and in determining subjective information like emotions and opinions (Zhang and Zheng 2016; Rumelli et al. 2019; Rathi et al. 2018). They surpass the traditional techniques in terms of cost, time-saving, and efficiency (Ayata et al. 2017; Abd El-Jawad et al. 2018; Zahoor and Rohilla 2020). Furthermore, the multinomial naïve Bayes technique is regarded as one of the most optimal machine learning algorithms since it is lightweight and scalable (Zhang 2004). It performs well on classification tasks and converges quickly even on multidimensional data because it is based on the conditional independence assumption in which the features are independent of one another (Zhang 2004). These characteristics are appropriate for Twitter data since the latter grows consistently over time. Furthermore, social media platforms host a large amount of data related to politics in general, and to the U.S. elections in particular, and it is almost impossible to process the data manually and infer insights from them to determine people's stances (Abd El-Jawad et al. 2018). In addition, social media data are valuable resources and are available for free. It would be inefficient to just ignore and get rid of them, and not to take full advantage of them to get useful insights and information. Furthermore, manual processing and on-ground public polls for retrieving people's opinions are very tedious and expensive tasks that need to be repeated completely for every election. However, automated software-based techniques such as sentiment analysis and machine learning techniques are a more convenient alternative in terms of cost, time, and efficiency.

The entire manuscript is organized as follows. First, a background, about the U.S. presidential election and the sentiment analysis approaches, is presented. The recent related works are then briefly discussed. The proposed model is introduced in the methodology section followed by results and discussion. Finally, the conclusion section summarizes the proposed work and discusses the research implications and future works.

2 Background

This section presents a brief background about the main two concepts used in this study: U.S. presidential election system and the geo-localization in the sentiment analysis.

2.1 U.S. presidential election system

The Electoral College is a process established at the founding of the country (Erikson et al. 2020; Kimberling 1992). While it was important for President to be elected by the people, it was recognized that a formal vote of electoral representatives might be required to settle disputes or if no candidate will win the majority of the votes. Each state has a certain number of electors (electoral college votes) based on population. The number of state representatives in Congress (435 in the House and 100 in the Senate) determines the number of Electoral College votes. Three more are allocated to the District of Columbia for a total of 538 Electoral College votes (Erikson et al. 2020; Kimberling 1992). For most states, the party that gets the most votes earns all the electoral college votes allocated to that state. The exceptions are Maine and Nebraska, which allocate Electoral College votes proportional to the vote in the state (Dunne 2012). The President requires a majority (270) of electoral votes to be elected (Erikson et al. 2020).

The U.S. presidential election system is mostly bipartite. Political parties nominate a president and vice-president who campaign to win the most votes in each state (Erikson et al. 2020). While other parties nominate candidates, the contest has mainly been between the Republican and Democratic parties. For the 2020 U.S. presidential election, Donald Trump was the nominee of the Republican party, while Joe Biden was the nominee of the Democratic party. Since the U.S. presidential election is a state-based system, a candidate can lose the total popular votes but still win if he wins the total electoral votes. This is what happened to George Bush in 2000 (Kriner and Reeves 2014; Vaughn 2013).

2.2 Geo-localization in the sentiment analysis

Sentiment analysis is a computational process in which a text is classified into three categories: neutral, positive, and negative. It is considered an opinion mining in which the stance of a user is determined (Bharti and Malhotra 2020). Geo-localization in sentiment analysis refers to identifying the sentiment of the user based on its geolocation. Geolocation information from social media especially Twitter can be leveraged for conducting geolocation-based analyses (Mousset et al. 2020; Zhong et al. 2020; Qarabash and Qarabash 2020; Han et al. 2014) such as election analysis (Jurgens et al. 2015).

Subsequently, a detailed analysis based on the location (i.e., state) of electorates is required for a better understanding of the public stance to predict accurately the election's outcome. Previous works (Zahoor and Rohilla 2020; Abroms and Lefebvre 2009; Nausheen and Begum 2018; Heredia et al. 2018; Avello et al. 2011; Metaxas et al. 2011) in the literature have failed to tailor a forecasting model that simulates the state-based system of the U.S. presidential elections which highly depends on the geolocation of electorates. This paper presents a model that simulates the U.S. presidential elections based on the electoral system that uses states as electoral entities. This model infers insights from social media outlets, more specifically Twitter, to predict the public stance of electorates in each state. It performs sentiment analysis on tweets while taking into consideration the geolocation of Twitter's users, which allows us to predict the public stance for each state, the winner of electoral votes, and hence the winner of the U.S. presidential elections.

3 Literature review

Researchers have been applying sentiment analysis on tweets (Calderon et al. 2015; Ferrara and Yang 2015), covering many areas ranging from studying the sentiments of customers from reviews (Asur and Huberman 2010) to predicting election results (Borondo et al. 2012). Recently, they started to focus more on the political aspect of sentiment analysis, such as examining Twitter's influence as a source of mainstream news (Morstatter et al. 2013; O'Connor et al. 2010; Parmelee 2013), studying Twitter's role during the election campaign (Abroms and Lefebvre 2009), analyzing the sentiment of citizens about government (Calderon et al. 2015), and predicting election outcome (Borondo et al. 2012). They even studied the congressmen's activities on Twitter during their campaigns (Glassman et al. 2010; Golbeck et al. 2010).

Although this proliferation in election prediction research, there is still doubt about how much this prediction simulates and matches the on-ground scenario (Avello et al. 2011; Metaxas et al. 2011) since a lot of demographic information is missing in Twitter data. However, many researchers still claim that they can predict accurate election results (Oikonomou and Tjortjis 2018; Tumasjan et al. 2010). A study on the Spanish elections shows that Twitter can be used as a tool for predicting election results in which they have developed a tool called "Taratweet,". This study shows that Twitter can be used for political discussions and that the mentions of the political parties and the electors' votes correlate significatively (Soler et al. 2012). This shows that political parties with a strong presence on Twitter during election campaigns receive more votes in the election. Another study on the French presidential elections shows that social network analysis and text mining techniques can be used for predicting election results. After scraping tweets related to the French presidential elections, this study shows how trends change affect people's opinions on Twitter and how they correlate with polling statistics (Wegrzyn-Wolska and Bougueroua 2012). Another research on the German federal election shows that Twitter can be used as a medium for expressing political opinions and as a credible indicator of political sentiment (Tumasjan et al. 2010). They studied how tweets reflect the on-ground political sentiment. They analyzed the political sentiment of 100,000 tweets affiliated with politicians and political parties and showed how it matches the election result. They also showed how joint mentions of parties match the on-ground political alliances. In 2016, a paper on the U.S. presidential election 2016 studied the people's sentiment in North Carolina, Florida, and Ohio using tweets related to two candidates: Hillary Clinton and Donald Trump (Oikonomou and Tjortjis 2018). They compared their results with the election results and showed how Twitter could predict the winner of the presidency.

Further studies on the U.S. presidential election cover the spatial factor of the sentiment analysis for predicting election results using Twitter data. In 2018, researchers published a paper discussing the efficacy of using location-based Twitter sentiment analysis to predict the U.S. 2016 presidential election (Heredia et al. 2018). They collected three million tweets related to Hillary Clinton and Donald Trump from September 22nd to November 8th for 21 states. They used tweets' sentiment and volume for predicting the election results. They conclude that their approach fails to reflect the actual election results. In 2020, location-based sentiment analysis of Twitter election data on the U.S. presidential election of 2016 and the UK general elections of 2017 evaluated the sentiment of location-based tweets with the election results (Yaqub et al. 2020). The U.S. election performs sentiment analysis on people's tweets related to two candidates: Hillary Clinton and Donald Trump for the following ten states: California, Florida, Georgia, Illinois, Michigan, New York, North Carolina, Ohio, Pennsylvania, and Texas. Another study in 2020 proposes an integrative model that substitutes poll data used in political science forecasting models with Twitter-based sentiments by using the 2016 U.S. presidential election in Georgia as a case study (Liu et al. 2021).

Most of the previous related works failed to find a model that simulates the U.S. presidential election system well. This manuscript proposes an efficient model that predicts the 2020 U.S. presidential election from geo-located tweets by leveraging the sentiment analysis potential, multinomial naive Bayes classifier, and machine learning. An extensive study on predicting the 2020 U.S. presidential election results is performed in which the state-based public stance for electoral votes and the public stance for popular votes are determined. The proper public stance is preserved by eliminating all outliers and removing suspicious tweets generated by bots and agents recruited for election manipulation. This work also studies the pre-election and post-election public stances and how they change in each state for each period. The influencers' effect on the public stance is also discussed. The effectiveness of the proposed model in predicting the election results is proved by comparing the close correspondence between the predicted outcomes and the actual election results. This work introduces a stance meter decision rule algorithm that generates the winner of the presidency. To detect any hidden patterns, network analysis, and community detection techniques are performed as well.

4 Methodology

Many previous works (Zahoor and Rohilla 2020; Abroms and Lefebvre 2009; Nausheen and Begum 2018; Rathi et al. 2018) leverage sentiment analysis techniques on social media for predicting election results. However, these works failed to account for the importance of the geolocation of Tweets, further possibly influencing the election outcomes. In nearly every state, the candidate who gets the most general votes wins the "electoral votes" for that state. Within 50 states, the magic 270 is the path to the presidency. This research project presents a geolocation-based multiclass classifier with a lexicon and rule-based sentiment analysis to classify tweets based on states and sentiments and classify users' stances toward each candidate.

The adopted workflow consists of four stages: acquisition, preprocessing (cleaning and vectorization/ tokenization), processing, and visualization. Figure 1 illustrates the proposed research model.

4.1 Data acquisition

The main aim of this manuscript is to analyze the public stance toward the candidates of the 2020 U.S. presidential election that was held on November 3rd, 2020,

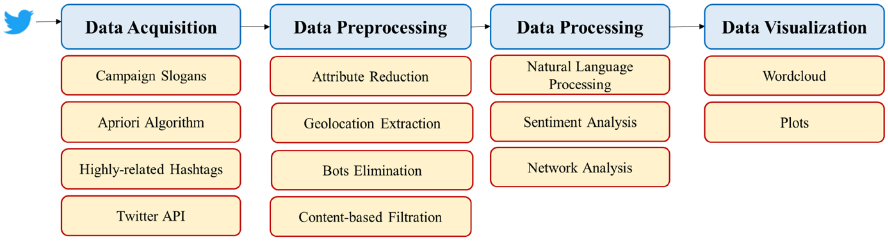

Fig. 1 The adopted workflow consisting of four stages: acquisition, preprocessing, processing, and visualization

and predict its outcome using Twitter data. To guarantee the completeness and closeness of the public stance, Twitter data were retrieved for a total of 25 days from October 15th to November 8th.

To better perceive the public stance shift before and after the election for both presidential candidates, the observed period is divided into two timeframes: the pre-election timeframe starting from October 15th, 2020 until election day (the end of November 3rd,2020), and the post-election timeframe from November 4th, 2020 till the end of November 8th, 2020.

The tweets were collected during the observed period using the Twitter API with election-related keywords such as #donaldtrump, #trump, #maga, #republican, #conservative, #makeamericagreatagain, #trumptrain, #keepamericagreat, #presidenttrump, #potus, #americafirst, #trumpsupporters, #gop as keywords for the first presidential candidate @realDonaldTrump and #joebiden, #biden, #democrats, #kamaharris, #voteblue, #bidenharris, #bluewave for the second candidate @JoeBiden, as they were the two primary presidential candidates.

Hashtags were extracted using the Apriori algorithm from slogans that were advocated by both candidates during their campaigns and then manually selected for both sides from the most frequently used ones for supporting and opposing the corresponding candidates.

After manually identifying the pro-Trump and pro-Biden hashtags from the highly related hashtags, 1,747,805 tweets from 483,212 users were collected, which is considered sufficient for the analysis. Figure 2 summarizes the data acquisition process.

4.2 Data preprocessing

Data preprocessing is a key for making the following tasks less challenging, thereby making the analysis takes less time. This process is divided into four main parts: attribute reduction, geolocation extraction, bots' elimination, and content-based filtration.

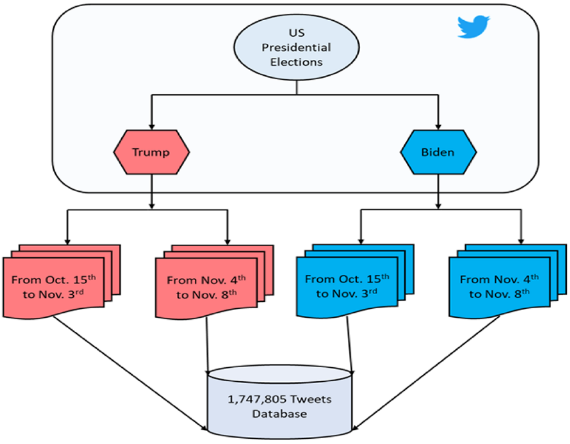
US Presidential Elections
Trump
Biden
From Oct. 15th to Nov. 3rd
From Nov. 4th to Nov. 8th
From Oct. 15th to Nov. 3rd
From Nov. 4th to Nov. 8th
1,747,805 Tweets Database

Fig. 2 Data acquisition of 1,747,805 tweets containing hashtags related to the U.S. presidential candidates Trump and Biden covering the pre-election (October 15th—November 3rd) and the post-election (November 4th—November 8th) timeframes

For every tweet, Twitter API yields a large bulk of metadata that contains irrelevant attributes and information. By that means, only the attributes that are relevant to the analysis are kept and summarized in Table 1.

Furthermore, tweets are geo-located by parsing location information such as latitude, longitude, city, country, state, and state code from 'user_location' and GPS information if available. To mitigate missing state locations, a geolocation dictionary is created to map latitude, longitude, and cities in their respective states. After extracting the location information of the users, tweets are subsequently categorized according to user state to be consistent with the U.S. presidential election system (electoral college).

In addition, the prediction should not only be based on the tweet per se but also on the tweet creator. So, it is noteworthy to conduct a bot check before processing the tweets to remove anomalous users. Bots and fake accounts that spread fake news and misleading information may affect the public stance. That's why, we eliminate users that don't show human behavior such as users with high activity, high posting rate, and joining date within one month span of the election day. In this context, we define these users as outliers that need to be investigated before considering them suspicious. For example, posting 2400 tweets in a 24 h period is considered a suspicious behavior whose user is marked as an outlier and hence eliminated from the analysis. Another example is a user who just joined Twitter a few days before the election and then did not exhibit any activity after the election period; this user is also marked as an outlier.

The aim of the content-based filtration is to retrieve valuable tokens from the text and get rid of the data that are semantically irrelevant. These data don't add much value to the meaning of the tweet and hinder the sentiment analysis. They include Twitter handles, hashtags, emoticons, URLs, special characters, newline characters, numbers, trailing whitespaces, and punctuation. After cleaning the tweets, tokenization is performed on the cleaned tweets, which is a pivotal step in the natural language processing pipeline. So, tweets are divided into a list sequence of tokens (bag of words) from which stopwords are removed. The remaining tokens are then stemmed by trimming suffixes such as "s," "es," and "ing."

4.3 Data processing

The data is divided into three main parts: natural language processing, sentiment analysis, and network analysis.

After cleaning and preprocessing the data, unigrams and bigrams are generated from the tokens, and their frequencies are calculated to get some contextual information about the tweets for natural language processing. The next step is to perform a sentiment analysis on the cleaned tweets and determine what insights can be obtained. So, to determine the public stance of Twitter users with respect to each presidential candidate while considering location, polarity analysis is used to quantify the sentiment of the tweets categorized into states.

One of the main challenges in applying sentiment analysis to social media content is its dynamic change. To mitigate this problem, a geolocation-based hybrid model is proposed.

Table 1 Relevant data attributes and their description

Data Attributes	Description
created_at	Tweet creation date and time
tweet_id	Tweet unique ID
tweet	Tweet text
likes	Number of likes
retweet_count	Number of retweets
user_id	Unique ID of tweet creator
user_name	Username of tweet creator
user_join_date	Date of when the user joined Twitter
user_followers_count	Followers count on user
user_location	Location is given on the user's profile

It consists of a Multinomial Naive Bayes machine learning classifier and a sentiment lexicon for the rule-based sentiment. The intuition behind this approach is to account for the multi-stance characteristics of tweets that are dual positive–negative paradoxes. For example, the same tweet on a topic can be negative for one candidate while positive for another. To solve these negative–positive Tweets, special tokens are generated to be biased more toward the subject than positive–negative bases.

For the Multinomial Naive Bayes machine learning classifier, a mathematical model is defined with a sentiment class $C_s \in \{\text{negative, neutral, positive}\}$ with $s=1$, 2, 3, and a feature vector v_i with $i=1, 2, \ldots, n$. The classifier is used to find the probabilities of sentiment classes assigned to tweets by using the joint probabilities of tweets' words and classes as expressed in Eq. 1.

$$P(C_s|v_1, ..., v_n) = \frac{P(v_1, ..., v_n|C_s)P(C_s)}{P(v_1, ..., v_n)} \tag{1}$$

For a given C_s, since each feature v_i is conditionally independent of every other feature v_j for $i \neq j$ as depicted in Eq. 2, Eq. 1 can be reduced to Eq. 3 which in turn can be expressed as in Eq. 4 since $P\ (v_1, \ldots, v_n)$ is constant.

$$P(v_i|C_s, v_1, ..., v_n) = P(v_i|C_s) \tag{2}$$

$$P(C_s|v_1, ..., v_n) = \frac{\prod_{i=1}^{i=n} P(v_i|C_s)P(C_s)}{P(v_1, ..., v_n)} \tag{3}$$

$$P(C_s|v_1, ..., v_n) \propto \prod_{i=1}^{i=n} P(v_i|C_s)P(C_s) \tag{4}$$

The model's classification rule is depicted in Eq. 5 and can be expressed in terms of log probabilities as Eq. 6 where $P\ (C_s)$ is the relative frequency of class C_s and $P\ (v_i \mid C_s)$ is the multinomial probability distribution of feature i belonging to the class C_s with $i=1, 2, \ldots, n$ where n is the number of features.

$$\hat{y} = \underset{s}{argmax} P(\prod_{i=1}^{i=n} P(v_i|C_s)P(C_s)) \tag{5}$$

$$\hat{y} = \underset{s}{argmax}(\sum_{i=1}^{i=n} lnP(v_i|C_s) + lnP(C_s)) \tag{6}$$

For each class C_s, the multinomial probability distribution $P\ (v_i \mid C_s)$ can be estimated by a smoothed version of maximum likelihood as expressed in Eq. 7 with α as the smoothing prior, N_i the number of occurrences of feature i belonging to class C_s, and N is the total number of occurrences of all features for class C_s.

$$\hat{y} = \underset{s}{argmax}(\sum_{i=1}^{i=n} ln\frac{N_i + \propto}{N + \propto n} + lnP(C_s)) \tag{7}$$

So, the sentiment score is computed in two rounds. For the first round, a Multinomial Naive Bayes classification approach is used for that purpose, whereas a lexicon and rule-based sentiment analysis approach is used for the second one as described in Algorithm 1. A sentiment score is returned from each round and assigned to tweets for both candidates. After normalizing these scores and computing their average, the mean sentiment score ranging from -1 to 1 is computed, which is used to classify each tweet as negative, neutral, or positive, where -1 represents absolute negative sentiment, 0 a neutral sentiment, whereas 1 represents an absolute positive one. Then, the geo-localization of the score is applied to cover the 50 U.S. states by adding a spatial attribute to the aforementioned scores. Subsequently, the mean sentiment for both presidential candidates is calculated for every state.

Algorithm 1 Proposed Method

```
Define $C$: Multinomial Naive Bayes Classifier
Define $L$: Lexicon and Rule-based Sentiment Analysis
Set $States$ = 1...50
Set $Candidates$ = 1...2

// Presidential candidates: Joe Biden and Donald Trump
for $c$ in $Candidates$ do
    //Geo-localization of the 50 US states
    for $s$ in $States$ do
        $round1 \leftarrow C$
        $round2 \leftarrow L$
        $score \leftarrow normalize\&average(round1, round2)$
        if $-1 \leq score < 0$ then
            $sentiment$ : "$Negative$"
        else if $0 < score \leq 1$ then
            $sentiment$ : "$Positive$"
        else
            $sentiment$ : "$Neutral$"
        end if
    end for
end for
```

Moreover, a social network of users interacting with each other and tweeting about the 2020 U.S. presidential elections is created to get some insights into the characterization of their interaction. A social interaction network $N\ (V, E)$ is defined with a set of vertices $v_i \in V$ representing users and a set of edges $e_{ij} \in E$ representing the interaction between two users v_i and v_j. To filter out relations between U users, the only considered edges are those with probabilities $p_{ij} < p_0$ where p_{ij} (Eq. 8) represents the probability of interaction between two corresponding users v_i and v_j with occurrence degree u_i and u_j, respectively, and a threshold probability $p_0 = 10^{-4}$. For each edge e_{ij}, a weight w_{ij} quantifying the linkage l between two users v_i and v_j is assigned according to Eq. 9. Furthermore, to identify community leaders and find their orientation, a community detection algorithm is applied with users as nodes and retweets as the edges between them.

$$p_{ij}(l) = \prod_{u=0}^{u=u_j-l-1} \left(1 - \frac{u_i}{U-u}\right) \prod_{u=0}^{u=l-1} \frac{(u_i-u)(u_j-u)}{(l-u)(U-u_j+l-u)} \tag{8}$$

$$w_{ij} = log(\frac{p_0}{p_{ij}}) \tag{9}$$

5 Results and discussion

In this section, an extensive study on predicting the 2020 U.S. presidential election results is performed. It answers all the following research questions:

Q1. How does the public stance vary for each presidential candidate based on the popular votes?
Q2. How does the state-based public stance differ based on the electoral votes?
Q3. How can a general stance meter decision rule algorithm be formulated from the experimental findings?
Q4. How does the public stance vary from the pre-election to the post-election with space and time?
Q5. How do influencers affect the public stance?
Q6. How can fraud, bot activity, and election manipulation be detected?
Q7. How can an electoral prediction model be depicted through network analysis and community detection?
Q8. Do the experimental findings reflect accurately the on-ground public opinion?

To predict the public stance for popular votes toward the two candidates: Donald Trump and Joe Biden, the context should be studied first. Figure 3 depicts the Bigram and Trigram for both candidates. The results show that the top ten are related to the two candidates and to the election, which is reasonably normal.

A WordCloud is a visualization that shows the most frequent words with a larger size than the less frequent words that appear smaller. It is used to get insights into the people's opinions about the two candidates and the most common words found

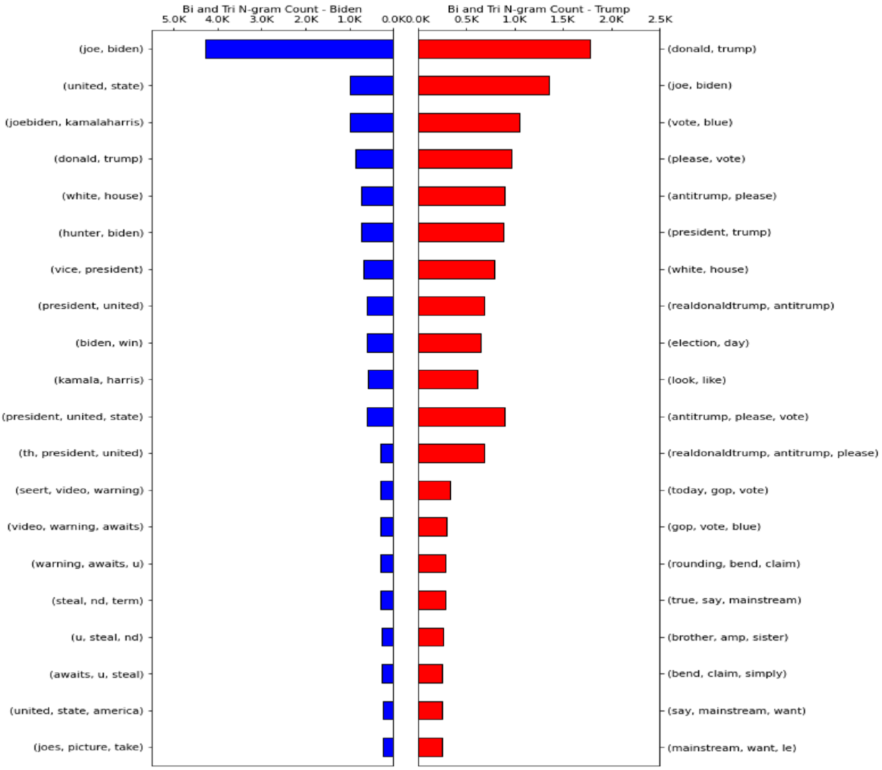

Fig.3 Bigram and Trigram for the two candidates Joe Biden and Donald Trump

in tweets. Figure 4 depicts the results of the WordClouds. The generated words are reasonable since they are related to the 2020 U.S. presidential election.

Figure 5 shows the density distribution of the sentiment score for both candidates in which the sentiment score ranges from -1 to + 1. In Fig. 5a, the peak is mainly concentrated and tends toward the positive region. This means that many people have positive sentiments toward Joe Biden. On the other hand, the peak of Fig. 5b is mostly concentrated and tends toward the negative region. This means that many people have negative sentiments toward Donald Trump.

To predict the public stance for popular votes, the percentage distribution of the sentiment score for both candidates is computed. The pie charts depicted in Fig. 6 illustrate the percentage distribution of the sentiment score for both candidates. Joe Biden is getting more positive sentiment than negative sentiment, whereas Donald Trump is getting more negative sentiment compared to positive sentiment. As a result, Joe Biden has more positive sentiment than Donald Trump despite the vast number of followers that Donald Trump has. We can conclude that Joe Biden wins in the popular votes.

The U.S. presidential election is based on the electoral college by its state-based system. The two candidates Joe Biden and Donald Trump, represent the Democratic

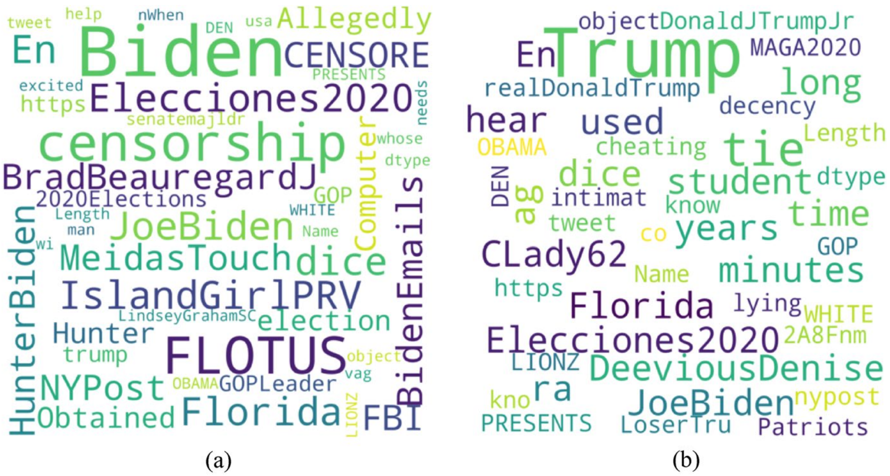

Fig. 4 The word clouds for the two candidates: **a** Joe Biden and **b** Donald Trump

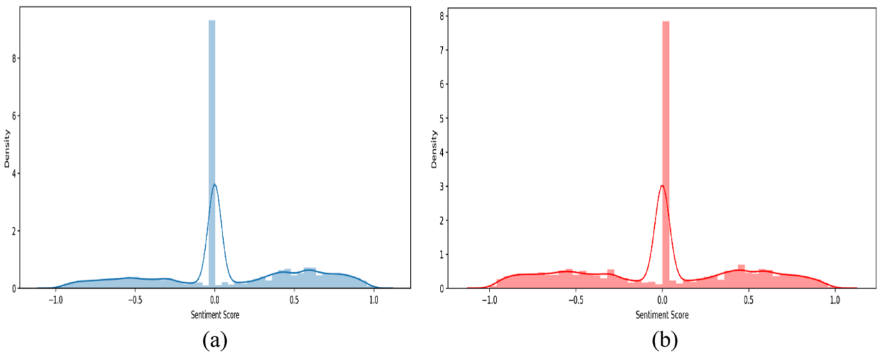

Fig. 5 The plot of the density distribution of the sentiment score for **a** Joe Biden and **b** Donald Trump

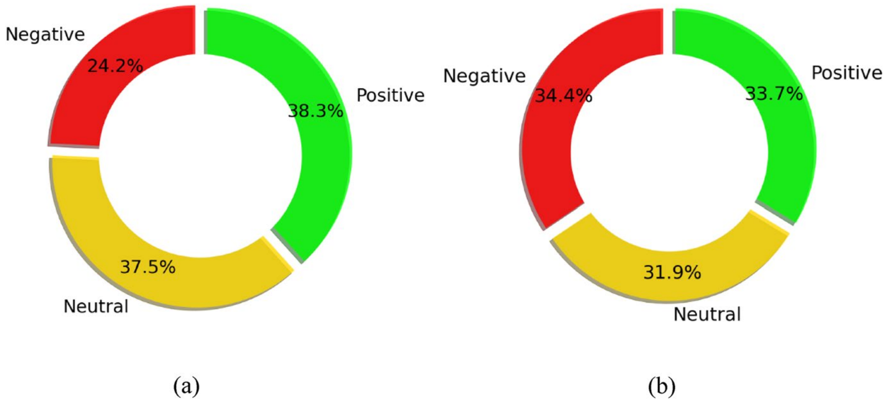

Fig. 6 The percentage distribution of the sentiment score for **a** Joe Biden and **b** Donald Trump

and Republican parties, respectively, and whoever wins most states wins the presidential election. Therefore, the sentiment score is studied for each candidate across the 50 U.S. states to know who the winner of the electoral votes is. Figure 7 shows the sentiment score for the 50 U.S. states for both candidates in which the higher the score, the more positive the sentiment is. Similarly, the lower the score, the more negative the sentiment is. If the score falls in the range between − 0.05 and + 0.05, the sentiment will be considered "weak". The weak zone is delimited in Fig. 7 by the two horizontal gray lines. For example, the sentiment is considered a "weak positive" if the score falls between 0 and + 0.05, whereas it is counted as a "weak negative" if the score falls between -0.05 and 0. The results show that most states are trending to a "strong positive" sentiment score for Biden and "weak negative" for Trump. Figure 8 shows the stance meter for both candidates across all 50 U.S. states. The stance meter classifies the stance based on the sentiment score of the tweets for each state. It generates the overall stance for each state represented by four output classes, including strong Republican, weak Republican, weak Democratic, and strong Republican. A stance score of + 1 corresponds to a "strong Republican" stance, which is mostly positive toward Trump, and that happens when the number of "positive" tweets related to Trump is greater than that of Biden, and the number of "negative" tweets related to Trump is smaller than that of Biden. Similarly, a stance score of −1 corresponds to a "strong Democratic" stance, which is mostly positive toward Biden, and that happens when the number of "positive" tweets related to Biden is greater than that of Trump, and the number of "negative" tweets related to Biden is smaller than that of Trump. Moreover, a stance score of + 0.5

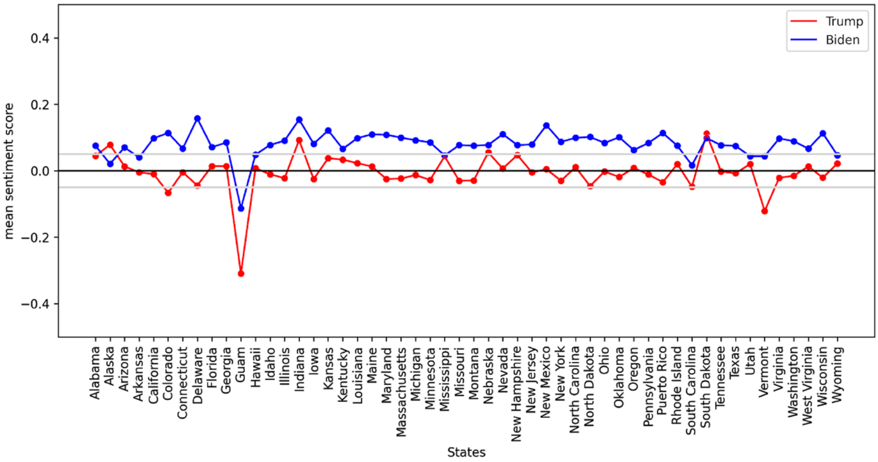

Fig. 7 The mean sentiment score for the 50 U.S. states for both candidates. The weak zone that ranges from − 0.05 and + 0.05 is delimited by the two horizontal gray lines. If the score falls within this zone, the sentiment will be considered "weak". A "weak positive" sentiment corresponds to a score between 0 and + 0.05 (upper gray line), whereas a "weak negative" sentiment corresponds to a score between − 0.05 (lower gray line) and 0. The higher the score, the more positive the sentiment is. Similarly, the lower the score, the more negative the sentiment is

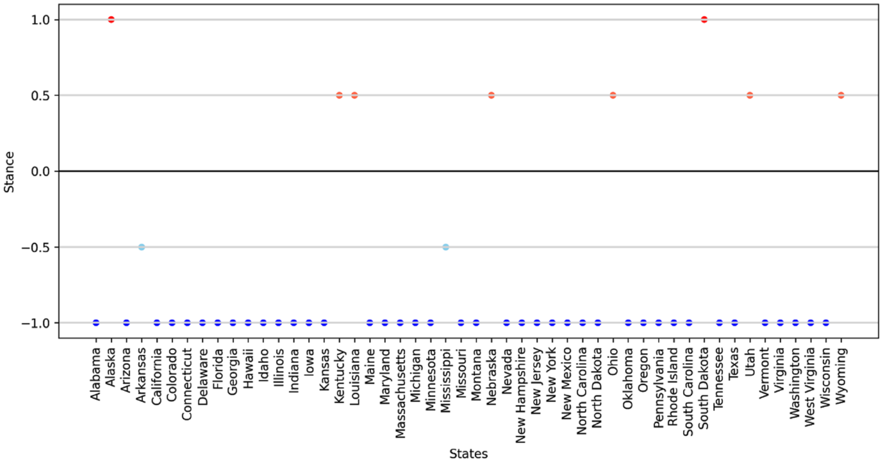

Fig. 8 The stance meter for both candidates across all 50 U.S. states. The Y-axis represents the overall stance for each state. The stance scores "1", "0.5", "– 0.5", and "– 1" correspond to the "strong Republican", "weak Republican", "weak Democratic", and "strong Democratic" stances, respectively. The red dots represent Trump (the Republican party), whereas the blue dots represent Biden (Democratic party)

corresponds to a "weak Republican" stance, and that happens when the difference between the number of "positive" tweets related to Trump, and the ones related to Biden is greater than the difference between the number of "negative" tweets related to Biden and the ones related to Trump. Similarly, a stance score of -0.5 corresponds to a "weak Democratic" stance, and that happens when the difference between the number of "positive" tweets related to Biden and the ones related to Trump is greater than the difference between the number of "negative" tweets related to Trump and the ones related to Biden. The red dots represent Donald Trump (the Republican party), whereas the blue dots represent Joe Biden (Democratic party). Since Joe Biden possesses more blue dots than his rival, he wins the majority of the electoral votes across all 50 U.S. states. According to the proposed model, Joe Biden wins the 2020 U.S. presidential election.

It is also interesting to study how the public stance varies with time for two periods: pre-election and post-election. Figure 9 shows how the positive, neutral, and negative stances vary for the two periods, pre-election and post-election. The specific dates where fluctuations occurred coincide with major political events such as the presidential debates. From the pre-election to the post-election phase, the positive and negative stances vary in favor of Joe Biden. The positive stance is increasing while the negative stance is decreasing.

It is also interesting to investigate the ten most active users for both candidates. Figure 10 shows the distribution of the ten most active users for Joe Biden and Donald Trump, in which Steve Ziegenbusch is the most active user from Biden's side, whereas Scott McLeod is the most active user from Trump's side.

In the U.S. presidential election, key influencers affect the public stance and make the people biased toward specific candidates. Since these influencers play an essential role in the U.S. presidential election, it would be great to investigate the effect of

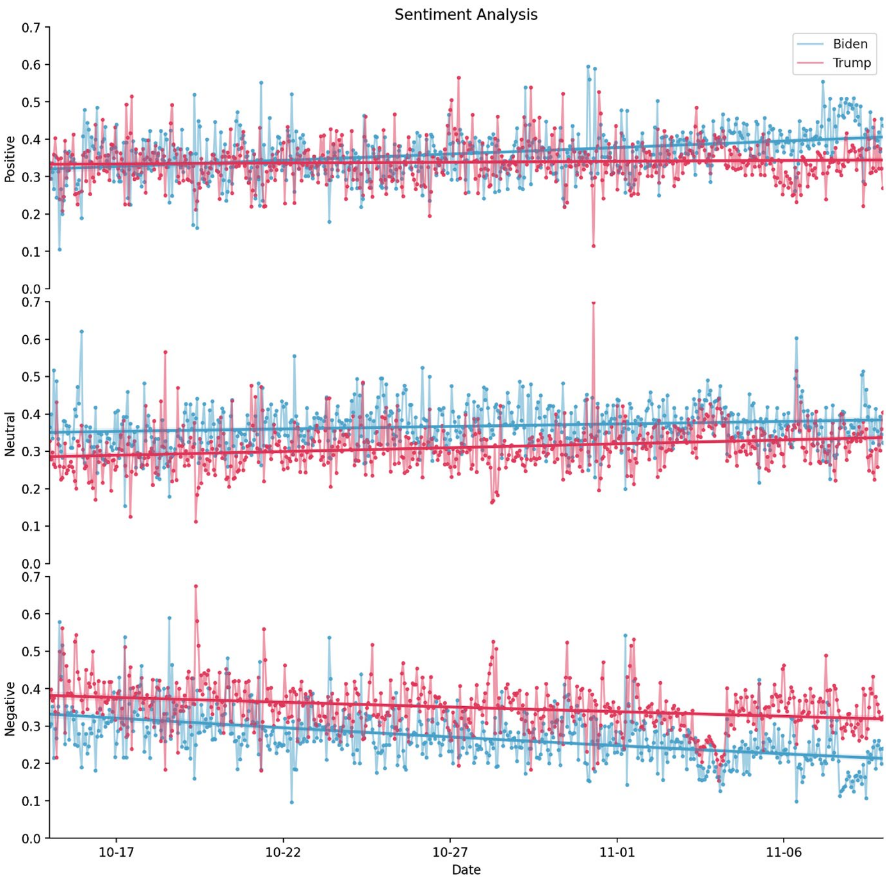

Fig. 9 The temporal change of positive, neutral, and negative stances for the two periods, pre-election and post-election

these influencers on the U.S. public stance. Figure 11 shows that Lady Gaga is the most influential in the U.S. presidential election. Knowing that she has a large popular base, she plays a pivotal role in supporting Joe Biden and promoting him, hence biasing the public stance toward him, thus affecting the election result in his favor.

Regarding the tweets' activities of the users, Fig. 12 shows the supporters' activity of both candidates and how it varies with time for both periods, pre-election and post-election. It also shows how the supporters react to the political events that have occurred. Mostly, the peaks occur on dates that correspond to the presidential debates and election day. More specifically, a spike occurs on October 23rd, which corresponds to the day following the final presidential debate that happened in Nashville, Tennessee. Furthermore, the supporters' activity of both candidates increases during the election day (November 3rd) and onwards, until it reaches its highest level on November 8th, which corresponds to the day that most national media organizations have projected Biden to win the 2020 U.S. presidential election.

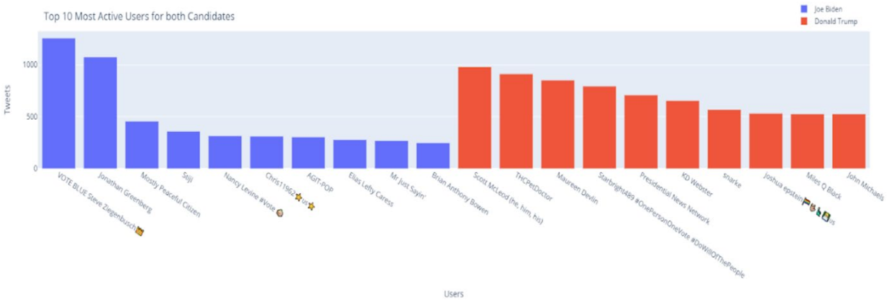

Fig. 10 The distribution of the most ten active users for both candidates

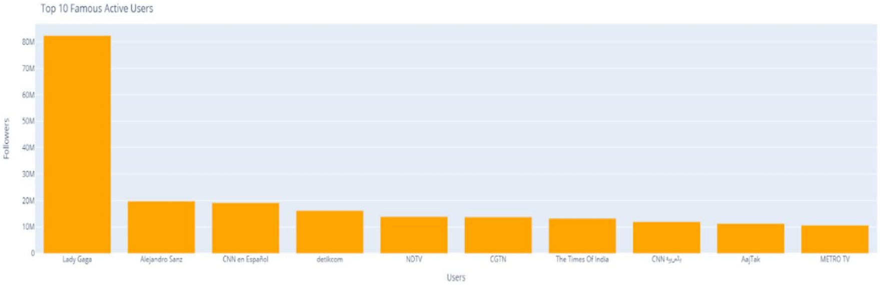

Fig. 11 The distribution of the most famous active users for both candidates

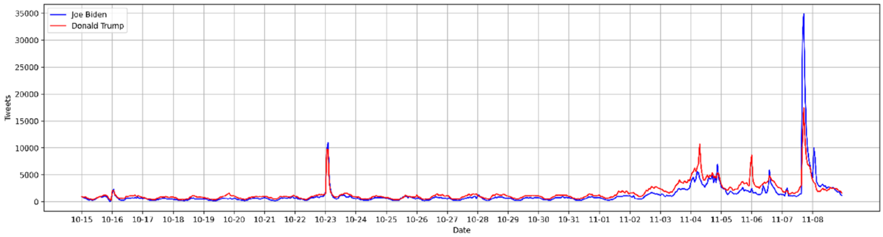

Fig. 12 The temporal variation of supporters' activity for both candidates. Spikes occur on dates that correspond to the day following the final presidential debate (October 23rd) and the period that follows the election day (November 3rd) with the highest level on the day of Biden projection (November 8th)

The supporters use Twitter as a platform for expressing their satisfaction with the election results.

We are also interested in investigating any possible manipulation in the election by detecting any suspicious users that try to manipulate the public stance and shift it toward a specific candidate. Figure 13 shows that the number of Trump supporters who have joined Twitter surpasses that of Biden's supporters. During most political events, there is a rise in the number of users who joined Twitter. This is an indicator of possible election manipulation. Moreover, all plots for both candidates show

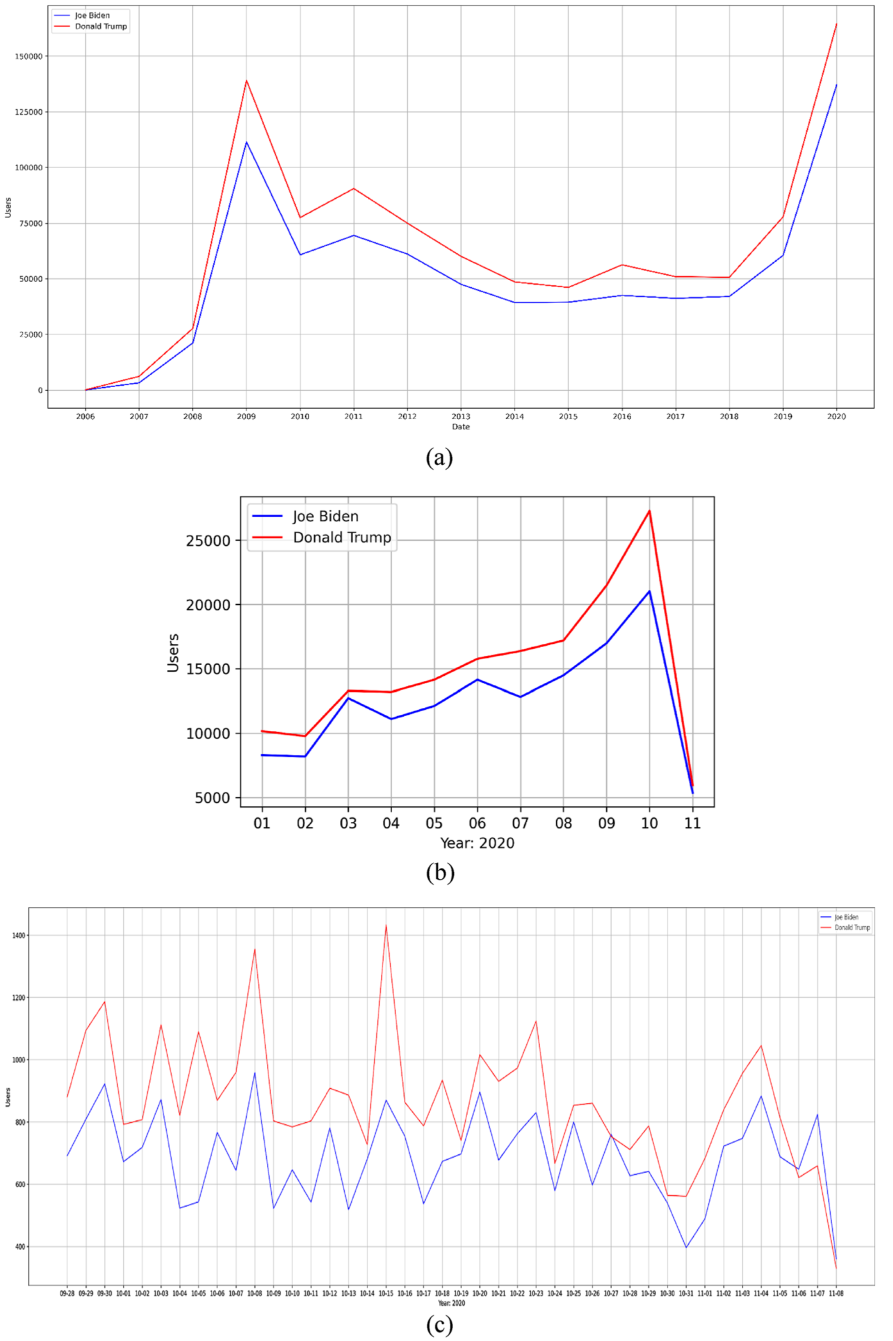

Fig. 13 The distribution of both candidates' supporters' joining date for: **a** all the years, **b** the year 2020, and **c** the electoral campaign period

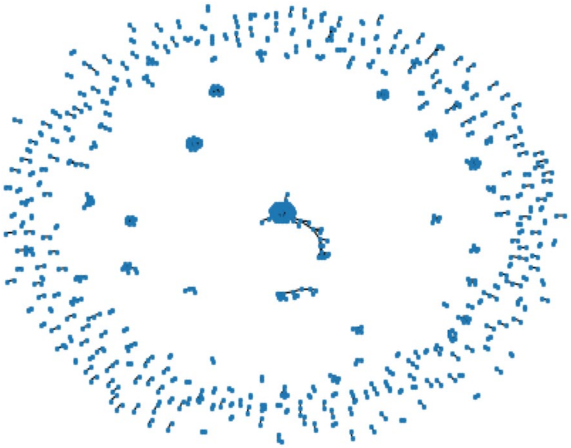

Fig. 14 The social network of the 2020 U.S. presidential elections with users as nodes and retweets as the edges between them

almost similar patterns, which might indicate fraudulent behavior. Figure 13.a shows that many users joined Twitter in 2020, which is the year of the presidential election. A possible reason for this behavior is tweeting about the election and attempting to manipulate the public stance toward a specific candidate or promoting his campaign. Figure 13.b illustrates the increase in Twitter users during 2020 to reach the maximum in October, just one month ahead of the election. This is probably due to the intention of biasing the public stance toward a specific candidate. Figure 13.c shows peaks that occurred during the presidential debates and election day. This can mean that users who have joined Twitter during these political events, especially election day, use Twitter to affect the public stance toward a specific candidate or spread fake news.

After performing the network analysis, the interaction between users tweeting about the 2020 U.S. presidential elections is illustrated in Fig. 14 as a social network N (V, E) which is constructed with a set of vertices $v_i \in V$ representing users and a set of edges $e_{ij} \in E$ representing the interaction between two users v_i and v_j.

Figure 14 identifies the community leaders from the social network which are the most important nodes in this network since they appear to be the center of the network and link most of the retweets. After extracting and investigating the nodes, most of these nodes are identified as Trump supporters, which makes them the most active community on Twitter and the most entities to have a high impact in promoting their candidate.

To simulate the electoral college system, the number of electoral votes assigned to each state should be considered in interpreting the stance meter results shown in Fig. 8. Table 2 shows the comparison between the predicted and the actual results of the U.S. presidential election with the corresponding electoral votes for each state. There are a total of 535 electoral votes (excluding the District of Columbia) out of which the majority goes to the Democratic candidate Joe Biden with a percentage of 89.9% despite the slight mismatch in the results, especially in some of the swing and battleground states. This outcome makes Joe Biden the winner of the electoral college, thereby the winner of the U.S. presidential election. This slight mismatch

Table 2 Comparison between the predicted and the actual results

State	Predicted Results	Actual Results	Electoral Votes
Alabama	Democratic	Republican	9
Alaska	Republican	Republican	3
Arizona	Democratic	Democratic	11
Arkansas	Democratic	Republican	6
California	Democratic	Democratic	55
Colorado	Democratic	Democratic	9
Connecticut	Democratic	Democratic	7
Delaware	Democratic	Democratic	3
Florida	Democratic	Republican	29
Georgia	Democratic	Democratic	16
Hawaii	Democratic	Democratic	4
Idaho	Democratic	Republican	4
Illinois	Democratic	Democratic	20
Indiana	Democratic	Republican	11
Iowa	Democratic	Republican	6
Kansas	Democratic	Republican	6
Kentucky	Republican	Republican	8
Louisiana	Republican	Republican	8
Maine	Democratic	Democratic	4
Maryland	Democratic	Democratic	10
Massachusetts	Democratic	Democratic	11
Michigan	Democratic	Democratic	16
Minnesota	Democratic	Democratic	10
Mississippi	Democratic	Republican	6
Missouri	Democratic	Republican	10
Montana	Democratic	Republican	3
Nebraska	Republican	Republican	5
Nevada	Democratic	Democratic	6
New Hampshire	Democratic	Democratic	4
New Jersey	Democratic	Democratic	14
New York	Democratic	Democratic	29
New Mexico	Democratic	Democratic	5
North Carolina	Democratic	Republican	15
North Dakota	Democratic	Republican	3
Ohio	Republican	Republican	18
Oklahoma	Democratic	Republican	7
Oregon	Democratic	Democratic	7
Pennsylvania	Democratic	Democratic	20
Rhode Island	Democratic	Democratic	4
South Carolina	Democratic	Republican	9
South Dakota	Republican	Republican	3
Tennessee	Democratic	Republican	11

Table 2 (continued)

State	Predicted Results	Actual Results	Electoral Votes
Texas	Democratic	Republican	38
Utah	Republican	Republican	6
Vermont	Democratic	Democratic	3
Virginia	Democratic	Democratic	13
Washington	Democratic	Democratic	12
West Virginia	Democratic	Republican	5
Wisconsin	Democratic	Democratic	10
Wyoming	Republican	Republican	3

that might make the stance meter appear inaccurate in terms of state-level outcomes is because the margin of victory in most of these states was under 5% in the actual election results, making them misclassified easily, thereby affecting the interpretation of the results.

All the research questions listed at the beginning of this section have been answered throughout this section. To sum up, the proposed model predicts Joe Biden as the winner of popular votes, satisfying the first research question (Q1). It also predicts him as the winner of electoral votes, satisfying Q2. To answer Q3, a stance meter is introduced and used as a decision rule algorithm that considers the electoral votes for each state. This stance meter was able to show that Joe Biden wins most of the electoral votes across all the U.S. states, making him the winner of the 2020 U.S. presidential election. After studying the pre-election and the post-election public stances and their variation with space and time, the model predicts that the public stance shifts toward Joe Biden as moving from pre-election to post-election, satisfying Q4. After studying the influencers' effect on the public stance, we found that Lady Gaga plays a significant role in supporting Joe Biden and answering Q5. The proposed model detected fraudulent behavior during political events such as the presidential debates and the election day, especially from Trump's supporters, answering Q6. As for answering Q7, the proposed model detected community leaders from the social network that illustrates the interaction between users tweeting about the 2020 U.S. presidential elections. Most of these community leaders are Trump supporters who are the most active community on Twitter and have a high impact in promoting their candidate. Finally, to answer Q8, the predictions match the actual presidential election results announcing Joe Biden as the winner of the 2020 U.S. presidential election.

For a better evaluation, an extensive comprehensive study is performed to cover all the aspects of the U.S. presidential elections as listed in the first column of Table 3. A comparative analysis with the most closely related works in the literature (Heredia et al. 2018; Yaqub et al. 2020, Liu et al. 2021) that cover the U.S. presidential elections is also performed, and the results are reported in Table 3.

The listed related works (Heredia et al. 2018; Yaqub et al. 2020, Liu et al. 2021) failed to cover most of the U.S. presidential elections' aspects listed in the

Table 3 Comparative analysis of the proposed work with the literature

	Heredia et al. (2018)	Yaqub et al. (2020)	Liu et al. (2021)	The proposed work
Election year	2016	2016	2016	2020
Number of states covered	21	10	1 (Georgia)	50
Method used	Deep learning	Lexicon	Machine learning	Hybrid*
Study period	Pre-election	Pre-election	Pre-election	Pre-election + Post-election
Electoral college system	No	No	No	Yes
Bot detection and elimination	No	Yes	No	Yes
Fraud detection	No	No	No	Yes
Influencers' effect	No	No	No	Yes
Network analysis and community Detection	No	No	No	Yes
General PUBLIC STANCE for popular votes	Yes	Yes	Yes	Yes
State-based public stance for popular votes	Yes	Yes	Yes	Yes
General public stance for electoral votes	No	No	No	Yes
State-based public stance for electoral votes	No	No	No	Yes
Stance variation	Spatial	Spatial + temporal	Spatial	Spatial + temporal
Comparison between the predicted and the actual results	Yes	Yes	Yes	Yes
Similarity with on-ground public opinion reflected in election results	Mismatching	Matching	Matching	Matching

*Multinomial Naive Bayes machine learning classifier + Lexicon and rule-based sentiment analysis first column of Table 3, more importantly, to simulate the electoral college system which is the core of the U.S. presidential elections. On the other hand, the proposed work which uses a hybrid method of classification for sentiment analysis is more tailored for the state-based electoral college system, especially in predicting the electoral votes' results. It incorporates both spatial and temporal variations from pre-election to post-election periods for all the U.S. states. Thus, it can be leveraged for future use in electoral campaigns as a complementary tool for forecasting the results of the U.S. presidential elections.

However, our work has potential limitations. The analysis is based on tweets that are written only in English, even though Spanish is the second most spoken language in the USA, and it is used predominantly on social media (Honeycutt and Sears 2020; Fuller 2012; Sheng et al. 2013). Moreover, the geolocation feature that is used in the proposed model is only limited to the state level. To better highlight the spatial influence on people's stance, further investigation is needed to include city and county levels in the analysis. These aspects are missing from our current study but will be addressed in our future works. Furthermore, many characteristics are also not considered in the analysis, including gender, race, and age. These characteristics can have an important impact on voting patterns. Therefore, they will be also part of our future works.

6 Conclusion and future works

Existing related works failed to find a model that resembles well the U.S. presidential election system and covers all the states. This manuscript proposes an efficient model that predicts the 2020 U.S. presidential election from geo-located tweets by leveraging the sentiment analysis potential, multinomial naive Bayes classifier, and machine learning. It answers all the research questions listed in Section 5. An extensive study about predicting the 2020 U.S. presidential election results is performed in which the state-based public stance for electoral votes is predicted. A stance meter is introduced and used as a decision rule algorithm that considers the electoral votes for each state. This stance meter was able to show that Joe Biden wins most of the electoral votes across all the U.S. states. The public stance for popular votes is also predicted. It projects that Joe Biden wins the popular votes across all the U.S. states. The true public stance is preserved by eliminating all outliers and removing suspicious tweets generated by bots and agents recruited for election manipulation. After examining the pre-election and the post-election public stances and their variation with space and time, the model predicts that the public stance shifts toward Joe Biden as moving from pre-election to post-election. The influencers' effect on the public stance is also discussed. It shows that Lady Gaga played a significant role in supporting Joe Biden. Network analysis and community detection techniques are also performed to detect any hidden patterns. The proposed model detected community leaders from the social network that illustrates the interaction between users tweeting about the 2020 U.S. presidential elections. Most of these community leaders are Trump supporters who are the most active community on Twitter and have a high impact in promoting their candidate. The effectiveness of the proposed model in predicting the election results is proved by comparing the close correspondence between the predicted outcomes and the real election results. The predictions match the actual presidential election results. A comparative analysis with the most closely related works in the literature is also performed to highlight the contributions and the novelty of our work. Our model outperforms all existing works and covers more aspects of the U.S. presidential election. The proposed model predicts Joe Biden as the winner of the popular votes and the electoral college with a percentage of 89.9%, making him the winner of the 2020 U.S. presidential election.

Our future works will address the limitations of the current study that are listed at the end of Section 5. The most spoken non-English languages in the USA will also be considered in the analysis, including Spanish, Mandarin, Tagalog, Vietnamese, Arabic, French, Korean, and Russian. This will investigate the effect of race and ethnicity on voting patterns, and how each ethnic group reacts to political events. Moreover, further analysis will target finer geolocation attributes, including city and county levels to better highlight the spatial influence on people's stance. Furthermore, many characteristics, including gender, race, and age, that can have an important impact on voting patterns, will also be considered in the future.

References

Abd El-Jawad MH, Hodhod R, Omar YMK (2018). Sentiment Analysis of Social Media Networks Using Machine Learning. 2018 14th International Computer Engineering Conference (ICENCO), pp. 174–176, Doi: https://doi.org/10.1109/ICENCO.2018.8636124.

Abroms LC, Lefebvre RC (2009) Obama's wired campaign: lessons for public health communication. J Health Commun 14(5):415–423

Asur S, Huberman BA (2010). Predicting the future with social media. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'10), Vol. 1. IEEE, 492–499.

Avello DG, Metaxas PT, Mustafaraj EN (2011). Limits of electoral predictions using Twitter. In Proceedings of the 5th International AAAI Conference on Weblogs and Social Media. Association for the Advancement of Artificial Intelligence.

Ayata D, Saraçlar M, Özgür A (2017). Turkish tweet sentiment analysis with word embedding and machine learning. 2017 25th Signal Processing and Communications Applications Conference (SIU), pp. 1–4, doi: https://doi.org/10.1109/SIU.2017.7960195.

Bharti O, Malhotra M (2020) Sentiment Analysis. SAGE Research Methods Foundations, Los Angeles

Borondo J, Morales AJ, Losada JC, Benito RM (2012) Characterizing and modeling an electoral campaign in the context of Twitter: 2011 Spanish presidential election as a case study. Chaos Interdisc J Nonlin Sci 22(2):023138

Calderon NA, Fisher B, Hemsley J, Ceskavich B, Jansen G, Marciano R, Lemieux VL (2015). Mixedinitiative social media analytics at the World Bank: Observations of citizen sentiment in Twitter data to explore "trust" of political actors and state institutions and its relationship to social protest. In Proceedings of the IEEE International Conference on Big Data (Big Data'15). IEEE, 1678–1687.

Dunne M (2012) The Long Winding Road to the White House: caucuses, primaries and national party conventions in the history of American presidential elections. Historian 115:6–12

Erikson RS, Sigman K, Yao L (2020) Electoral college bias and the 2020 presidential election. Proc Natl Acad Sci USA 117:27940–27944

Ferrara E, Yang Z (2015) Quantifying the effect of sentiment on information diffusion in social media. PeerJ Comput Sci 1:e26

Fuller JM (2012) Spanish Speakers in the USA (MM Textbooks, 9). Multilingual Matters

Glassman M, Straus JR, Shogan CJ (2010). Social networking and constituent communication: Member use of Twitter during a two-week period in the 111th Congress. 66.

Golbeck J, Grimes JM, Rogers A (2010) Twitter use by the U.S. Congress. J Amer Soc. Inf Sci Technol 61(8):1612–1621

Han B, Cook P, Baldwin T (2014) Text-based twitter user geolocation prediction. J Artif Intell Res 49:451–500

Heredia B, Prusa JD, Khoshgoftaar TM (2018). Location-based Twitter sentiment analysis for predicting the U.S. 2016 presidential election. In Thirty-First International Florida Artificial Intelligence Research Society Conference (FLAIRS-31), 2018.

Honeycutt L, Sears E (2020) Teaching Spanish in the United States in the digital age: strategies and approaches on teaching Spanish in online and hybrid classes. Bellaterra J Teach Learn Language Lit 13:838

Jurgens D, Finethy T, McCorriston J, Xu Y, Ruths D (2015). Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice. ICWSM.

Kimberling WC, National Clearinghouse on Election Administration (U.S.) (1992) The Electoral College. National Clearinghouse on Election Administration, Federal Election Commission, Washington, D.C

Kriner D, Reeves A (2014) The electoral college and presidential particularism. Bost Univ Law Rev 94(3):741–766

Liu R, Yao XA, Guo C, Wei X (2021) Can we forecast presidential election using twitter data? an integrative modelling approach. Ann GIS 27:43–56

Metaxas PT, Eni Mustafaraj E, and Dani Gayo-Avello DG. (2011). How (not) to predict elections. In Proceedings of the IEEE 3rd International Conference on Privacy, Security, Risk and Trust (PASSAT'11) and the IEEE 3rd International Conference on Social Computing (SocialCom'11). IEEE, 165–171.

Morstatter F, Jürgen Pfeffer J, Huan Liu H, and Kathleen M. Carley KM (. 2013). Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. In Proceedings of the International Conference on Web and Social Media (ICWSM'13).

Mousset P, Pitarch Y, Tamine L (2020) End-to-end neural matching for semantic location prediction of tweets. ACM Trans Inform Syst 39:1–35

Nausheen F, Begum SH (2018). Sentiment analysis to predict election results using Python. 2018 2nd International Conference on Inventive Systems and Control (ICISC), 1259–1262.

O'Connor B, Ramnath Balasubramanyan R, Bryan R. Routledge BR, and Noah A. Smith NA. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In Proceedings of the International Conference on Web and Social Media (ICWSM'10) 11, 122–129 (2010), 1–2.

Oikonomou L, and C. Tjortjis C (2018). A method for predicting the winner of the USA presidential elections using data extracted from Twitter. In 2018 South-Eastern European Design Automation, Computer Engineering, Computer Networks and Society Media Conference (SEEDAC ECNSM), pages1–8, Sep.2018.

Parmelee JH (2013) Political journalists and Twitter: Influences on norms and practices. J Media Pract 14(4):291–305

Qarabash NA, Qarabash H (2020) Twitter location-based data: evaluating the methods of data collection provided by twitter API. Int J Comput 19(4):583–589

Ramzan M, Mehta S, Annapoorna E (2017). Are tweets the real estimators of election results? 2017 Tenth International Conference on Contemporary Computing (IC3), pp. 1–4, doi: https://doi.org/10.1109/IC3.2017.8284309.

Rathi M, Malik A, Varshney D, Sharma R, Mendiratta S (2018). Sentiment Analysis of Tweets Using Machine Learning Approach. 2018 Eleventh International Conference on Contemporary Computing (IC3), pp. 1–3, doi: https://doi.org/10.1109/IC3.2018.8530517.

Rumelli M, Akkuş D, Kart Ö, Isik Z (2019). Sentiment Analysis in Turkish Text with Machine Learning Algorithms. 2019 Innovations in Intelligent Systems and Applications Conference (ASYU), pp. 1–5, doi: https://doi.org/10.1109/ASYU48272.2019.8946436.

Sayce D (2021). The number of tweets per day in 2020 | David Sayce. Retrieved February 16th, 2021, from https://www.dsayce.com/social-media/tweets-day/https://www.dsayce.com/social-media/tweets-day/

Sehl K (2021). Top Twitter Demographics That Matter to Social Media Marketers. (2021). Retrieved February 16th, 2021, from https://blog.hootsuite.com/twitter-demographics/https://blog.hootsuite.com/twitter-demographics/

Sheng L, Bedore LM, Peña ED, Fiestas C (2013) Semantic development in Spanish-English bilingual children: effects of age and language experience. Child Dev 84(3):1034–1045. https://doi.org/10.1111/cdev.12015

Shi L, Agarwal N, Agrawal A, Garg R, Spoelstra J (2012). Predicting U.S. primary elections with Twitter. Retrieved fromhttp://snap.stanford.edu/social2012/papers/shi.pdf.

Soler JM, Cuartero F, Roblizo M (2012).Twitter as a Tool for Predicting Elections Results. Proc. IEEE/ACM Int'l Conf. on Advances in Social Networks Analysis and Mining, Istanbul, pp. 1194–1200.

Tumasjan A, Sprenger TO, Sandner PG, Welpe IM (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment, Proc. 4th Int'l AAAI Conf. on Weblogs and Social Media, pp. 178–185, 2010.

Vaughn JS (2013) New perspectives on the public presidency: the impact of elections and presidential travel. Pres Stud Q 43(3):671–673

Wolska KW, Bougueroua L (2012). Tweets mining for French Presidential Election. Proc. 4th Int'l Conf. on Computational Aspects of Social Networks (CASoN). Sao Carlos, 2012, pp. 138-143.

Yaqub U, Sharma N, Pabreja R, Chun SA, Atluri V, Vaidya J (2020) Location-based sentiment analyses and visualization of twitter election data. Digit Gov Res Pract. https://doi.org/10.1145/3339909

Zahoor S, Rohilla R (2020). Twitter Sentiment Analysis Using Machine Learning Algorithms: A Case Study. 2020 International Conference on Advances in Computing, Communication & Materials (ICACCM), pp. 194–199, doi: https://doi.org/10.1109/ICACCM50413.2020.9213011.

Zhang H (2004). The Optimality of Naive Bayes. FLAIRS Conference.

Zhang X, Zheng X (2016). Comparison of Text Sentiment Analysis Based on Machine Learning. 2016 15th International Symposium on Parallel and Distributed Computing (ISPDC), pp. 230–233, doi: https://doi.org/10.1109/ISPDC.2016.39.

Zhong T, Wang T, Zhou F, Trajcevski G, Zhang K, Yang Y (2020). Interpreting Twitter User Geolocation. ACL.

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Dr. Rodrigue Rizk is currently a professor at the University of South Dakota. Dr. Rizk received the B.E. degree in computer and communication engineering Summa Cum Laude highest honor distinction from Notre Dame University and was the valedictorian of his class with a GPA of 4.0/4.0. He received his M.S. and Ph.D. degrees in Computer Engineering from the University of Louisiana at Lafayette while maintaining a perfect grade point average. His area of specialization is comprised of the dynamic relationship between software and hardware. His research interests include high-level computational systems, artificial intelligence, data science and analytics, Internet of Medical Things (IoMT), data privacy and security, reinforcement learning, quantum and neuromorphic computing, high-performance computer architecture, deep learning optimization, hardware-software co-design, heterogeneous computing, Field Programmable Gate Arrays (FPGA), VLSI, AI hardware accelerators, healthcare, epigraphy, ancient languages, and emerging technologies. Dr. Rizk has collaborated actively with researchers in several other disciplines of informatics, computer science, and engineering, ranging from theory to design to implementation, and has published several research papers in top-tier conferences and journals. Dr. Rizk is a licensed Professional Engineer with a wide range of industry expertise and a member of the Order of the Engineer in the United States. He is the recipient of the prestigious Richard G. and Mary B. Neiheisel endowed fellowship, a lifetime member of the Phi Kappa Phi honor society, and a professional member of ACM and IEEE. He is also a member of the IEEE Standards Association, IEEE Computer and Quantum Society, ACM-W, and Emerging Interest Groups (EIG) on Smart Connected Communities and Reproducibility and Independent Verification. Dr. Rizk is also the recipient of many prestigious awards including the President's Award for Educational Excellence and Outstanding Academic Achievement and the Ragin' Leadership Academy Award.

Dr. Dominick Rizk received the B.E. degree in Computer Communication Engineering Summa cum laude highest honor distinction from Notre Dame University, in 2017, and his M.S. and Ph.D. degrees in Computer Engineering with a cumulative GPA = 4.0/4.0 from the University of Louisiana at Lafayette, Louisiana, USA, in 2020 and 2023, respectively. His research interests mainly include IoT, reconfigurable design systems, hardware security with a concentration on physical unclonable functions (PUFs), machine learning-based modeling attacks on PUFs and countermeasures. He was awarded the University of Louisiana at Lafayette Dissertation Completion Fellowship. He is a lifetime member of the Phi Kappa Phi honor society.

Frederic Rizk received the B.E. degree in computer and communication engineering with Summa cum laude highest honor distinction from Notre Dame University in 2018. He received the M.S. degree in Computer Engineering from the Center of Advanced Computer Studies (CACS), University of Louisiana at Lafayette, USA in 2021, where he is currently a Ph.D. candidate and has maintained a perfect grade point average. His research interests lie in the areas of machine and deep learning with concentration on GAN, adversarial attacks, and countermeasures. He is a lifetime member of the Phi Kappa Phi honor society.

Dr. Sonya Hsu is an associate professor in the School of Computing and Informatics, University of Louisiana at Lafayette. While pursuing academic programs in Telecommunication at Michigan State University and then Ph.D. in Management Information System at Southern Illinois University at Carbondale, Dr. Hsu had rigorous research methodology and data analysis training. Combining data-centric enthusiasm and ERP knowledge, Dr. Hsu recently explores machine learning and deep learning. Both quantitative and qualitative work can be seen in Dr. Hsu's publications in the Journal of Business Research, Information and Management, and the Hawaii International Conference on System Sciences (HICSS). Her research focuses on information and data analytics of operational excellence and patient care.

Authors and Affiliations

Rodrigue Rizk[1] · **Dominick Rizk**[2] · **Frederic Rizk**[2] · **Sonya Hsu**[2]

Dominick Rizk
dominick.rizk1@louisiana.edu

Frederic Rizk
frederic.rizk1@louisiana.edu

Sonya Hsu
sonyahsu@louisiana.edu

[1] Department of Computer Science, University of South Dakota, Vermillion, SD 57069, USA

[2] Center for Advanced Computer Studies , University of Louisiana at Lafayette, Lafayette,
LA 70504, USA