USA Presidential Election Sentiment Analysis Using TextBlob on Real Time Twitter Data

Shaon Sikder1, Md Al-Imran2, Dr. Mohammad Arifuzzaman3

1,3Department of Electronics and Communications Engineering, East West University,  
2Department of Computer Science and Engineering, East West University,  
Aftabnagar, Dhaka 1212, Bangladesh   
E-mail:1shaonsikder.ewu@gmail.com, 22020-1-96-011@std.ewubd.edu, 3mazaman@ewubd.edu

**Abstract**

Social networking sites have become an indispensable part of the business and political campaigns. Therefore, it contains much information that can be used to predict market supply-demand or political popularity. People express their views on any issue on social media; those data, later on, can be utilized to understand the stance of the people on a particular issue. In this paper, we use real-time twitter data expressing views on issues of the USA Presidential Election. We explored the data to make the insight, and, then based on the insight, we applied the TextBlob library to find the polarity. What we have derived from the experiment, we utilize the outcome to interpret the data. Thus we reached an end commenting on both parties present situation. To complete this paper, we use the NLTK sentiment analysis package along with the TextBlob library of python.

*Keywords*: NLP, NLTK Corpus, TextBlob Sentiment Analysis, US Election 2020, Data Preprocessing

# Introduction

Social media has started dominating vital decisions of a community as well as a country. People express their perspectives, creating and sharing opinions on national or communal issues, suggesting or supporting ideas through posting as blogs or informative video messages on social media platforms. YouTube, Twitter, Facebook, Instagram, WhatsApp are some of the social sites which have been creating an impact on the vital phenomenon. For example, YouTube played a massive role behind the victory of the Democratic Party and Barack Obama as the President of the USA in 2008[1]. The Democratic Party victory increased the popularity of online campaigning from that time. Twitter is a microblogging media. The word tweet refers to delivering short expressions or messages regarding any issue. The tweet impacts the societal and economic structure as the USA election mentioned above[2]. D. Trump and H. Clinton utilized the social media platform for online campaigning during the United States Presidential Election of 2016. **#**MakeAmericaGreatAgain was the hashtag of Republican in the tweets at that time. Parties circulated sufficient funds to advertise their respective views towards the voters. In 2020, both parties have carried out the online campaign as well. Republicans and Democrats have been expressing their commitment to economic and national issues on the social platform to inform the world. As a result, citizens and non-citizens are reacting to the posts and expressing their perspectives on different policies enlisted in the manifesto. Since there are millions of tweets worldwide regarding the presidential election, we keep our focus on mining the tweets data. Inspired by data science research history of the previous presidential election, we applied sentiment analysis on tweets data. As mentioned above, social media is such a platform where user-generated contents are available such as- public opinion on products or receiving mass opinion through polls under different news journal. Specifically, twitter allows us to share, deliver, and interpret real-time tweets. In this paper, we have mined the real-time tweets which have given us insights into the present situation of two major presidential candidates of the USA. The sentiment analysis helps us to describe the popularity and possibility of one of the potential candidates to lead the next USA. Massive rich sets of data obtained from the twitter are analyzed in sentiment analysis.

It is well known, there are many approaches to sentiment analysis, such as- Linguistic Inquiry and Word Count (LIWC)[3] is used to extract features for texts. Usually Naive Bayes, SVM supervised learning algorithms are also used in sentiment analysis. A labeled training data set is required to perform a supervised learning algorithm. However, in this paper, we perform analysis within the political domain based on the twitter data collected according to some keywords given as the parameter. In this research, our focus to visualize the following- i) candidate popularity and engagement; ii) buzz and significant ideologies which creating impact; iii) opinion polarization; iv) visualizing real-time data plot; v) finding which candidate is leading in the race. It is now clear that the content of our research is bounded within the area of digital platform politics.

Additionally, a reasonable explanation may be drawn according to the findings. Since public opinion polarization is another concern to us, we would show how it is potentially impacting on the election. Our preferred method is a pre-trained model like TextBlob to perform the NLP research, which has simple API for sentiment analysis in Python programming language. The remaining sections are organized as follows. The section-2 summarizes the related article. Section-3 illustrated the methodology of collecting tweets and preprocessing of data. Section-4 contains model implementation and architecture. Section 5 visualizes the findings and discusses the results, then section 6 concludes.

# Related Works

In [4], authors proposed sentiment analysis based on the pre-processed data frame. The authors exploit TF-IDF model vectorizer to find improved accuracy. The author identifies the sentiment using positive and negative polarity. The author incorporates various NLP techniques like-lemmatization, stop word removal, parts of speech tagging, and TF-IDF model. In the last decade, Twitter data has been utilized to predict various domains like- business and economics, social issues, politics, and online campaigns etcetera.

In [5], authors provided a model to analyze real time public sentiment toward the 2012 US presidential candidates using a statistical approach to understand political practices through twitter users. An algorithm proposed by [6] can classify tweets as polarity with respect to asking questions. The authors use Naive Bayes classifier to identify the polarity as a class. Vast numbers of features make a model of higher variance; apparently more complicated model has higher variance in general. The authors use frequency based feature selection, Mutual Information, and CHI-Square based feature selection methods. The author uses the Maximum Entropy classifier, SVM, and Naive Bayes classifier and analyzes the performance. The effect of fake news on browsing is described in [7]. The authors estimate 10 article increases in fake news increase the odds of reading 3.7% news sites.

In [8], authors perform sentiment analysis of user tweets to find the correlation with real news and buzz issues in context of the US Election 2016. Then the author assesses the impact on tweets by the candidates since users took part in the debate. Depending on these, the author finally identifies twitter as a platform for political discussion and interaction.

A dataset with emoticons and hashtags is used by [9]. The authors use several techniques such as- parts of speech tagging, *n*-gram features, lexicon features, and microblogging features. The author focused on linguistic features analysis from the twitter messages. In [10], the author proposed a new entity level sentiment analysis method which adopts lexicon based sentiment analysis. The author uses SVM as a classifier to assign polarities to the tweets. Lexicon based approach is helpful for the non labeled data. So far, we have enlisted authors work basically gives an insight of how social media analytics investigates the nature of voter. In [2], the author shows the behaviour of the voter analyzing tweets through polarization of voter preferences. The study provides relevant insights into voter behavior for future political campaigns. Valence Aware Dictionary for sEntimenet Reasoner (VADER) is used to classify the sentiments in twitter data by the author [3]. The findings of the study based on the 2016 USA election showed satisfactory accuracy. The author uses NodeXL visualization and open-source network analysis. To work with the textual data in NLP NLTK provide an easy way to use interfaces of 50 corpora and lexical resources, tokenization, text processing libraries for classification, parsing, stemming, tagging, and semantic reasoning. VADER is a lexicon and rule based entirely free open-source sentiment analysis tool.

# Research Methodology

This section presents the hypothesis of our work and focuses on phases of sentiment analysis and result visualization. The work consists of five phases as shown in Fig. 1. The first phase concerns with fetching twitter data under some defined keywords. The second focus is based on dataset creation and preprocessing of data to remove irrelevant information from the tweets. Phase three deals with sentiment analysis model building with TextBlob and Natural language toolkit NLTK. In phase four, we applied the technique to analyze the data; then, the results are visualized with the seaborn python library in phase five. The study aims to evaluate the current assessment of the candidates of the USA Presidential Election 2020 among the twitter users based on the real time twitter data.

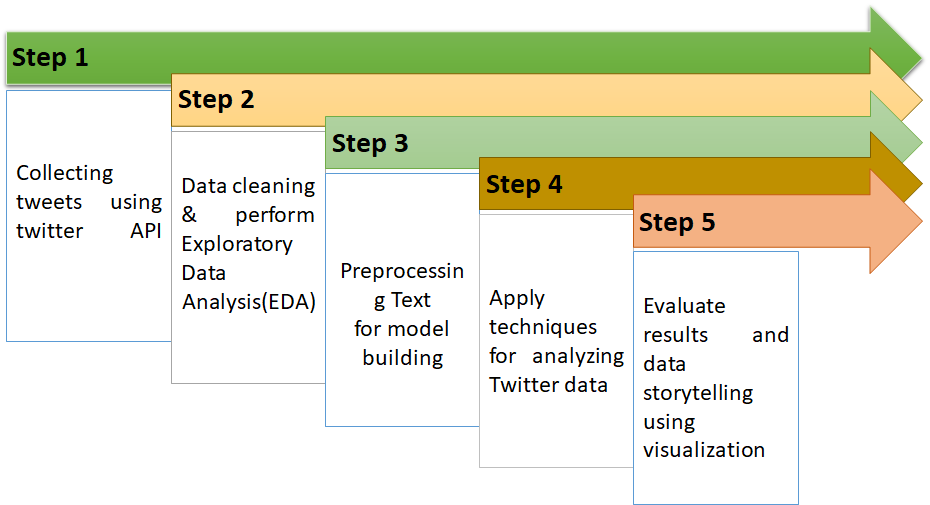


Fig. 1. Five Phases of the Study

## Data Set Creation

The two main competitors of the USA presidential election are Donald Trump and Joe Biden, and their running mates are Mike Pence and Kamala Harris. Twitter data for two candidates and their running mates were collected on July 22, 2020. We used the Twitter streaming API to fetch the data with the help of python tweepy library. The tweepy library provides access to the twitter streaming API quickly. We have collected the data by passing three keywords as input parameters to the respective streaming method in tweepy. The keywords are biden, kamala, democrat for democratic party data, and trump, mikepence, republican are for the republican party data. We separately collected data for each keyword; after collecting all the data, we append the data to create a dataset consisting only of the Democratic Party’s data; the process remains the same for the Republic Party dataset. The dataset is stored in a CSV file later on. In the following Table 1 and Table 2, we have shown the data count for the presidential candidates for two major parties.

Table 1

Data Count for Republican

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Keywords | trump | mikepence | republican | Total |
| Count | 2500 | 500 | 2500 | 5500 |

Table 2

Data Count for Democrat

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Keywords | biden | kamala | democrat | Total |
| Count | 2500 | 500 | 2500 | 5500 |

## Data Cleaning

Data cleaning is required to get better sentiment analysis. Firstly, we removed the duplicate and null rows from our dataset. Then all texts are converted into lowercase using the regular expression module of python. Do we need punctuations, numeric values, special characters, stop words and the text containing numeric values? Our objective of this work does not require to keep these items in our dataset. So we removed these items using re module of python.

## Corpus

In this portion, we will cover how we used the corpus of NLTK to train the model, such as a part-of-speech tagger or stop-words removing. Stopwords are those words that generally do not contribute to the principal meaning of sentences. For the purpose of information retrieval in NLP, we need to remove these words like – the, a, and, apart, anyhow etcetera. Removing stop words also save space in memory. In stop-word removing, there is a way of tokenization of the sentences and then applying the NLTK’s default method. Since our data is split before this stop, we look upon the NLTK corpus. NLTK has come with stopwords corpus that has lists of the word for many languages[11]. To avoid complexity, we have cleaned the data earlier to make it free from punctuation and unnecessary formats. It paves the way of making it easier to apply built in *nltk.corpus* module. The stop words corpus is an instance of *nltk.corpus*. It has a *words()* method that takes an argument. We have given ‘english’ as stop words list. The lowercase and split data set is mapped with those stop words; if any found, it is removed from the data frame.

## Exploratory Data Analysis

Suppose your friends want to watch a new movie with you. However, you do not know what the movie is about! At this stage what will you do? Definitely, you will ask about the cast and crew, producer and director. Furthermore, you might watch the trailer and find the rating and review of the movie at IMDB. Whatever assessment procedure you consider, in the end, it will give you confidence whether you spend time with the movie or not. Exploratory Data Analysis (EDA) is more or less like that.

Data scientist explores the data to understand its variety and tries to find out insights from the data. EDA is a philosophy to carry out data analysis that incorporates statistical techniques such as uncover underlying structure, anomaly and outlier detection, maximize insight into data.

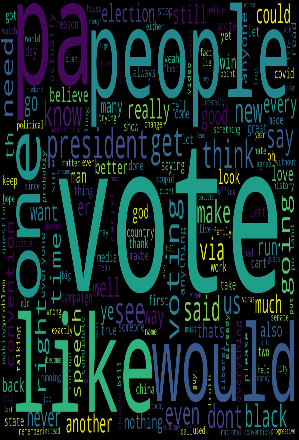
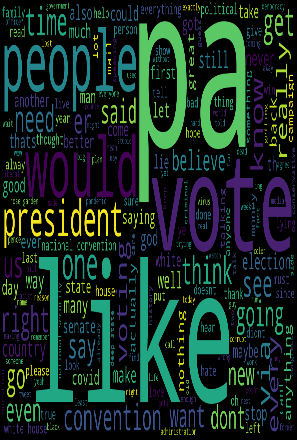
 

Fig. 2. Word Cloud for Democrat (left) and Republican (right)

At first glance at this step, we found that keywords in sentences are redundant, so we remove it using a regular expression module. Additionally, we look into the data description to find out the uniqueness of data. However, there are still some redundant sentences; in fact, sentences with less than one word are an outlier for our study. So we removed those outliers due to clean our data frame to make some insights. We used the word cloud to see whether more exploration is required or not. We visualize words in the word cloud in Fig. 2. Moreover, the most used words are providing some hints which will be discussed in the result section.

## Data Preprocessing

Data is fetched from twitter account of various users. Therefore, there should be data quality assessment. We cannot rely on those fetched data. There may be flaws in data that we need to deal with. The following is some of the problems we faced:

* Null Values: It is very much usual to have null values in your dataset. Null values are removed from the dataset.
* Duplicate values: Objects which are duplicates of one another might be dropped. Otherwise, the particular duplicate data object may give bias predictions.

# Sentiment Model Building

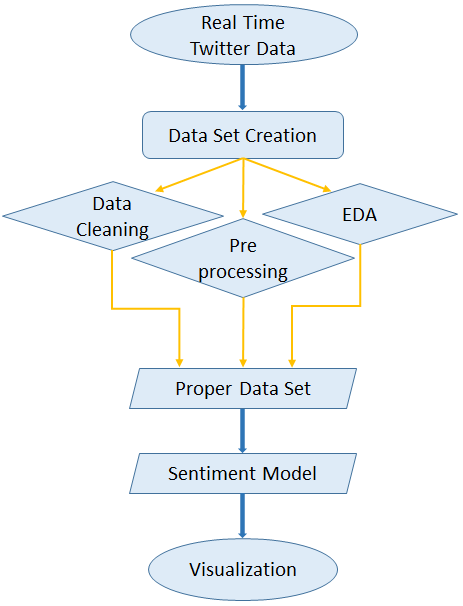


Fig. 3. The architecture of the Study

Now we have a dataset, and the dataset is preprocessed and cleaned. The design of the sentiment model used in our study is based on the expectation of getting the polarity of public opinions of twitter data. NLTK is such a package that provides tools for classifying data. NLTK has a feature that can provide the sentiment of each sentence. For our study, we applied the NLTK procedures to the dataset separately to identify whether the sentiment of a tweet is positive or negative or neutral concerning presidential candidates.

In Fig. 3, the architecture of our study has shown—the real time twitter data has been processed to create the outlier free dataset. Then the sentiment model has been deployed. Using TextBlob, we have defined a function that returns the polarity of each sentence such as- *sentiment.polarity* > 0 assign 1 to the sentence. A portion of pseudocode has been given here-

1. **analysis TextBlob(data)**
2. **if analysis.sentiment.polarity > 0:**
3. **return 1**
4. **else if analysis.sentiment.polarity is 0:**
5. **return 0**
6. **else:**
7. **return -1**

Polarity expresses the sentiment polarity of the tweet. The polarity is set between -1 to 1. Positive polarity tweets are classified as a high chance of occurrence, and negative polarity tweets are classified as the chance of not happening. From Table 3, we can observe that the sentiment is classified into three polarity- positive, neutral, and negative. In the following section, we are going to present the result and prediction based on visualization.

Table 3

Sentiment Analysis Polarity

|  |  |  |
| --- | --- | --- |
|  | Tweets | Sentiment |
| 4714 | black people still pushing trope problem white... | -1 |
| 4548 | light covid response owe apology guy wro rix | 1 |
| 4305 | yep definitely falling keep | 0 |

# Results

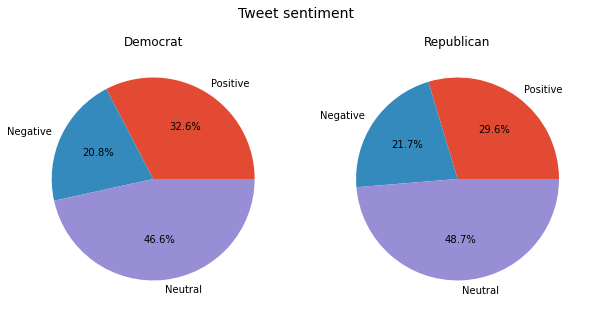


Fig. 4. Pie Chart showing percentage of polarity

TextBlob finds all of the words and phrases that it can assign a polarity and subjectivity to, and averages all of together. As a result, each row of the dataframe will assign with one polarity and subjectivity score, as mentioned in the previous section with an appropriate figure of output. To get a clear picture of the current wining position based on sentiment, we classified three polarities separately for each party. Furthermore, after that, we created a pie chart to provide an interpretation.



Fig. 5. Positive sentiment percentage of two parties

From Fig. 4, the democrat pie chart can be described as-

* There is 32.6% of the population who tweeted on any issues of the Democratic Party, showed a positive attitude towards the victory.
* 20.8% of the population turned down the policy of Democrat candidates.
* 46.6%, many of the population maintained neutrality, as they might think of better policy proposals yet to publish.

From Fig. 4, the republican pie chart can be described as-

* There is 29.6% of the population who have a firm belief in the current president.
* 21.7% of the population showed disagreement towards the ruling party.
* 48.7%, again, a large number of the population is hoping for better policy.

From these two pie charts, one point is exact, there is too much neutrality towards politics, which indicates a significant portion of the population lost their interest in participation due to the controversial policy of parties around different times. In Fig. 5, we can observe the twitter data analysis percentage for both parties. It is clearly seen that the Democratic Party is ahead of the Republican Party with a small margin.

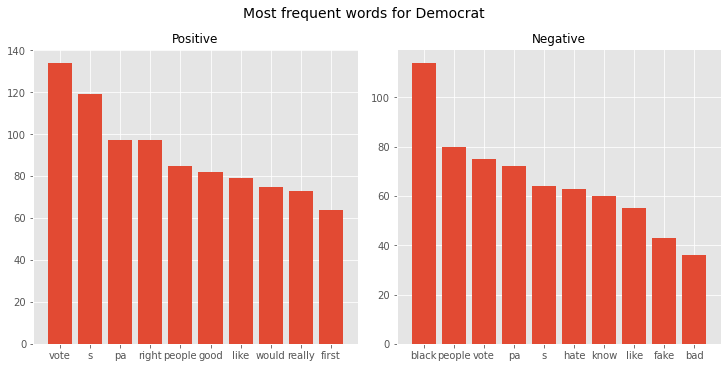


Fig. 6. Most frequent words for Democrat

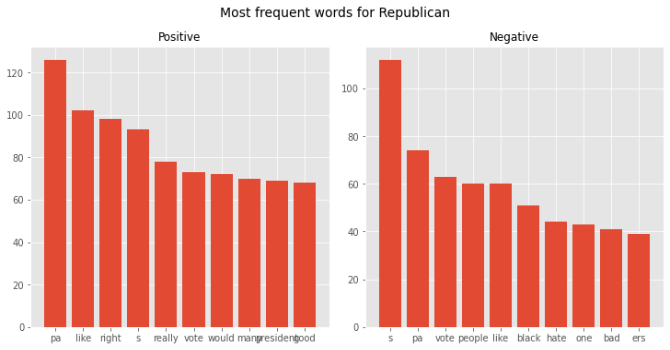


Fig. 7. Most frequent words for Republican

Fig. 6 shows the most frequent positive and negative words used for the Democratic Party. This shows that Democrat is spreading more kindness to the oppressed and deprived portion of population. The party should propose strict policies on the rights of the black community and immigrant rights. From Fig. 7, it is found that people are criticizing the hatred and deprivation of the black community under the current government. The top words for Republicans in terms of positive and negative sentiment were limited mostly to the black community and bad decisions assessed based on tweet.

# Conclusions

In this paper, we have developed a model for sentiment analysis that can be used anytime to make assumptions about the presidential election of the USA. This model can be used anywhere to analyze the data which is utilizing NLTK. To conclude, based on the result, it is easily observed that Republican Party have to propose better policy regarding the equal rights of all community keeping in mind that, the USA is not built by a particular community alone. Therefore, the world is really expecting some good policy from wining party which will be beneficial for the mankind.

**References**

[1] S. L. Stirland, “Propelled by internet, barack obama wins presidency,” *Wired Mag.*, vol. 4, 2008.

[2] P. Grover, A. K. Kar, Y. K. Dwivedi, and M. Janssen, “Polarization and acculturation in US Election 2016 outcomes--Can twitter analytics predict changes in voting preferences,” *Technol. Forecast. Soc. Change*, vol. 145, pp. 438–460, 2019.

[3] S. Elbagir and J. Yang, “Twitter Sentiment Analysis Using Natural Language Toolkit and VADER Sentiment,” in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, 2019, pp. 116–122.

[4] M. R. Hasan, M. Maliha, and M. Arifuzzaman, “Sentiment Analysis with NLP on Twitter Data,” in *2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)*, 2019, pp. 1–4.

[5] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, “A system for real-time twitter sentiment analysis of 2012 us presidential election cycle,” in *Proceedings of the ACL 2012 system demonstrations*, 2012, pp. 115–120.

[6] A. Go, L. Huang, and R. Bhayani, “Twitter sentiment analysis,” *Entropy*, vol. 17, p. 252, 2009.

[7] A. R. Doshi, S. Raghavan, R. Weiss, and E. Petitt, “How the Supply of Fake News Affected Consumer Behavior during the 2016 US Election,” *Available SSRN 3093397*, 2018.

[8] U. Yaqub, S. A. Chun, V. Atluri, and J. Vaidya, “Analysis of political discourse on twitter in the context of the 2016 US presidential elections,” *Gov. Inf. Q.*, vol. 34, no. 4, pp. 613–626, 2017.

[9] E. Kouloumpis, T. Wilson, and J. Moore, “Twitter sentiment analysis: The good the bad and the omg!,” in *Fifth International AAAI conference on weblogs and social media*, 2011.

[10] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, “Combining lexicon-based and learning-based methods for Twitter sentiment analysis,” *HP Lab. Tech. Rep. HPL-2011*, vol. 89, 2011.

[11] J. Perkins, *Python 3 text processing with NLTK 3 cookbook*. Packt Publishing Ltd, 2014.