"Sentiment Analysis on Twitter Data Regarding US Presidential Election 2020 (1498 words)"

Pratap Roy Choudhury, MS in Data Science, Spring 2022

Indiana University Bloomington

1. INTRODUCTION

The general election of 2020 in United States caused quite a discussion on social media. The sentiment analysis technique is one of the best ways to understand the public mood and opinion towards the candidates - Republican current president Donald Trump and Democratic challenger Joe Biden [2]. Twitter, one of the biggest platforms for microblogging has seen an influx of tweets from users [4][5]. A fraction of this tweets was organic, i.e., people who were genuinely expressing their opinions on social media regarding their choice of the leader. Also, political innuendo from various campaigns and slandering the oppositions using twitter as a platform was also used. Through analysis of this online discourse, we can discover the main tendencies and preference of the electorate, generate patters that can distinguish users' favoritism towards an ideology or a specific political party, study the sentiment prevailed towards the political parties or even predict the outcome of the elections [1]. This research will study about the reactions of the people who are in favor of Joe Biden becoming the president, and people who are opposing it.

2. RESEARCH QUESTION

The objective of this research paper is to analyze people's reactions and comments posted in Twitter regarding US presidential Election in 2020. The study will mainly address the following research question:

- What are the people's reactions and sentiment regarding Joe Biden running for the President of the United States of America?

3. METHODOLOGY

3.1. DATA COLLECTION

The data is collected using snscrape between October 25, 2020 - November 3, 2020. The Twitter data extraction process snscrape is a scraper for social networking services (SNS). It scrapes things like user profiles, hashtags, or searches and returns the discovered items, e.g., the relevant posts. As Tweepy is unable to extract historical data from Twitter, we are using the snscrape module to collect the tweets related to the 2020 US president election.

To keep the data distribution for each campaign uniform, the data collection is done in two sets - first using the hashtags - #Election2020, #JoeBidden for Joe Biden related tweets and secondly, using the hashtags - #DonaldTrump, #Trump along with some keywords - "2020 election", "Joe Biden", "Donald Trump", "US Election 2020". A python script is executed by feeding these hashtags and keywords as search parameters, the start date and end date between 2020-10-25 to 2020-11-03 and English as the query language. In this way, 101 tweets for each date are extracted and thus 1010 data points are available for each of the candidates - Joe Biden and Donald Trump.

Below are the details of each of the fields of the dataset:

Field	Description
Username	String that displays the twitter handle or user profile name used in Twitter
Datetime	Date when the tweet was posted (yyyy-mm-dd hh-mm-ss format)

Tweet ID	The unique tweet identifier posted by the user
Text	The tweet/comment posted by the user

Table 1: ***Data dictionary of the tweets collected using snscrape***

3.2. DATA ANALYSIS

To perform sentiment analysis on the textual data collected, the texts are needed to be cleaned to remove the unwanted characters and form a meaningful document on which a classification model can be applied. The complete analyses steps are explained below.

3.2.1. Text Preprocessing

In Natural Language Processing, the textual data is needed to be cleaned and processed before modeling. With the help of some manual observation, it is found that the tweets contain various characters, symbols, http links, emojis etc. which should be removed. The prerequisite steps are as follows -

I. The texts are transformed to lower-case and then any username (starting with @), # symbols, numbers, URLs (starting with https://), characters that are non-alphanumeric, 'rt' for the re-tweets are removed from the text using regular expression.

II. A list of stop-words is prepared specifically for this use case. Words such as not, did not, doesn't etc. are not removed as these words can impact the positive and negative sentiment of the text. All such stop-words and punctuations are removed from the text data.

This procedure removes the text noise and finally it allows the identification of the entity that was discussed by the users. Now the sentiment analysis can be done on the cleaned text.

3.2.2. Sentiment Analysis

In the implementation of sentiment analysis, VADER analysis model from python NLTK library is utilized [6]. This implementation technique is chosen because it is especially attuned to the sentiment expressed in social media and sensitive to both polarity (positive/negative) and intensity (strength) of emotion.

As there are two files of tweets - one for each candidate, the plan is to find the sentiment of the tweets for each group separately which will give an idea about the people's reaction for the individual candidate.

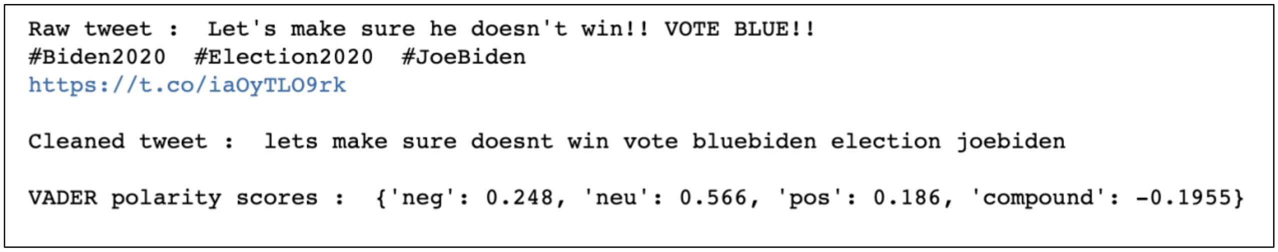
UserName	Datetime	Tweet Id	Text	candidate	cleaned_text
HelloKarma2021	2020-10-25 23:47:23+00:00	1.3205E+18	#JoeBiden is making a fool out of himself https://t.co/XUs2KcEn3n	Joe Biden	joebiden is making a fool out of himself
MartyR214	2020-10-25 23:43:55+00:00	1.3205E+18	@CortesSteve Real President #JoeBiden	Joe Biden	real president joebiden
JasonRFate	2020-10-25 23:18:11+00:00	1.3205E+18	Tweeted 1 year ago today #JoeBiden @JoeBiden #vote #2020Election https://t.co/iGjkVMeJ3m	Joe Biden	tweeted year ago today joebiden vote election

Table 2: ***Sample of tweet dataset for Joe Biden after text preprocessing***

UserName	Datetime	Tweet Id	Text	candidate	cleaned_text
PresidentRat	2020-10-25 23:59:16+00:00	1.3205E+18	This is not what our great 2nd Amendment. #Trump	Donald Trump	this is not what our great nd amendment trump
RohitMGaikwad	2020-11-03 23:56:21+00:00	1.3238E+18	@domjoly #DonaldTrump is winning again #USAElections2020	Donald Trump	donaldtrump is winning again usaelections
bit33dotio	2020-10-27 23:58:13+00:00	1.3212E+18	President #Trump's Campaign Website Was #Hacked https://t.co/ihHdtwKbgF #hackers #Election2020	Donald Trump	president trumps campaign website was hacked hackers election

Table 3: ***Sample of tweet dataset for Donald Trump after text preprocessing***

VADER sentiment analysis provides polarity_scores() method to every tweet text in order to understand the sentiment of the tweet. The result of this method call is a dictionary showing the intensity of negative, neutral, and positive sentiment in the tweet. All these three values are used to create the fourth figure which is the overall compound sentiment of the tweet.



```
Raw tweet :  Let's make sure he doesn't win!! VOTE BLUE!!
#Biden2020  #Election2020  #JoeBiden
https://t.co/iaOyTLO9rk

Cleaned tweet :  lets make sure doesnt win vote bluebiden election joebiden

VADER polarity scores :  {'neg': 0.248, 'neu': 0.566, 'pos': 0.186, 'compound': -0.1955}
```

Figure 1: ***Example of polarity scores of a cleaned text using VADER sentiment analysis***

In this analysis, two rule based algorithms are developed.

3.2.2.1. Algorithm 1:

By comparing the positive, negative, and neutral polarity scores, rules are generated to classify the sentiment of a text.

- if pos >= neg and pos >= neu, the sentiment = ‘Positive’
- if neu >= pos and neu >= neg, then sentiment = ‘Neutral’
- else sentiment = ‘Negative’

3.2.2.2. Algorithm 2:

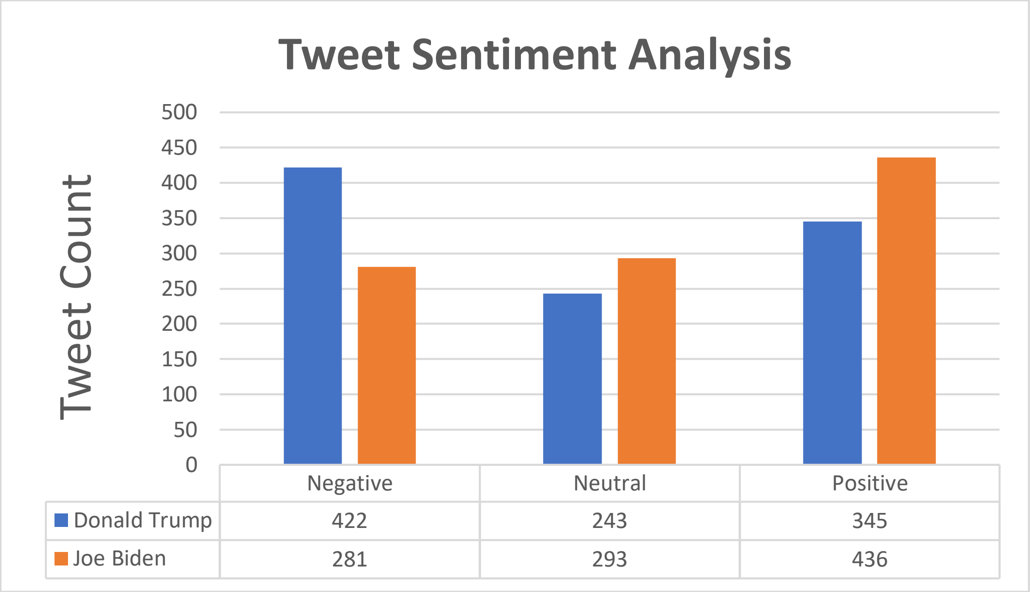
The compound score is basically the normalized sum of all the polarity scores where -1 being extreme negative and +1 being extreme positive. The probable rule can identify texts with compound score between -0.05 and +0.05 as a neutral sentiment. A quick manual observation of the compound scores and the actual sentiment of the texts helped to manipulate the compound score range to classify the sentiment of all the tweets.

- If compound >= 0.05, then sentiment = ‘Positive’
- If compound <= -0.1, then sentiment = ‘Negative’
- else sentiment = ‘Neutral’

The results from both the algorithms are compared for a set of tweets manually and the algorithm 2 is finalized for the classification. The results are mostly neutral in algorithm 1 which contradicts the actual sentiment of the tweets when evaluated manually. The algorithm 2 classifies the tweet sentiments almost accurately.

4. RESULTS

I. As there are 1010 data points for each candidate groups, it's found that most of the tweets are in support of Joe Biden with 43.17% positive, 27.82% negative and 29.01% neutral tweets. On the other hand, there are 41.78% negative, 34.16% positive and 24.06% neutral tweets for Trump. There are more positive tweets for joe Biden and more negative tweets for Donald Trump shared by the people. Figure 2 shows the tweet sentiment comparison for both the candidates.



	Negative	Neutral	Positive
🟦 Donald Trump	422	243	345
🟧 Joe Biden	281	293	436

Figure 2: ***Tweet sentiment distribution - out of 1010 tweets for each candidate***

II. An interesting insight that has been found from the sentiment trend lines from the dataset of each candidate (on 1010 individual data points, 101 tweets per day) is that the negative tweets were dropped significantly in the last 2-3 days for both the groups. The increase in positive tweets shows that as the election date approached, people advertised the candidate of their choice, rather than denouncing the opposing candidate.

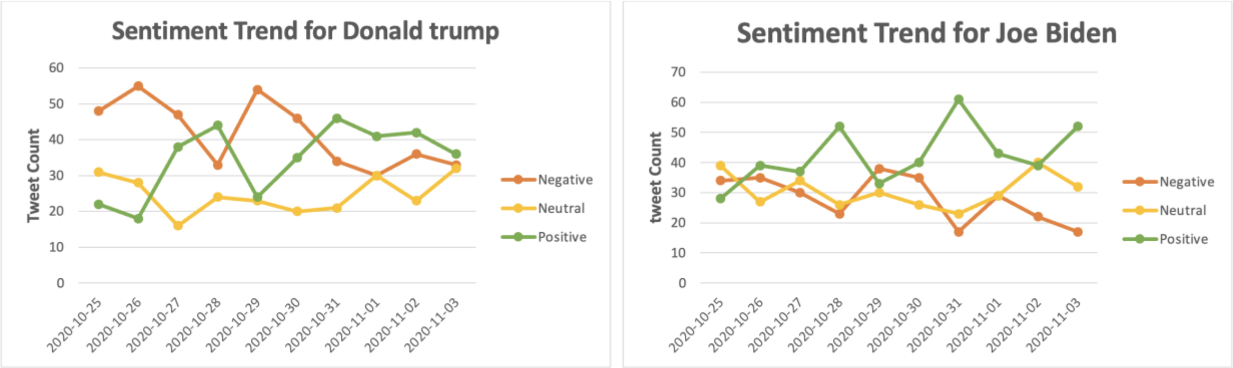


Figure 3: ***Predicted tweet sentiment trend for Trump and Biden between 2020-10-25 to 2020-11-03***

(*Trend is based on the sentiment counts of individual dataset of Joe Biden and Donald Trump*)

III. The tweets that are collected for Donald Trump, are also considered for tweet sentiment count as complementary data for Biden. In this analysis, it is found that the dataset for Trump also contain some keywords related to Joe Biden or the Democrats. So, the positive and negative tweets from Trump dataset are considered as negative and positive sentiment respectively for Joe Biden. The neutral tweets are counted as neutral. Thus there are 2020 data points for Joe Biden. Analysis shows that, Joe Biden had a great support of people as 42% of the tweets were positive, 31% were neutral and only 27% were negative.

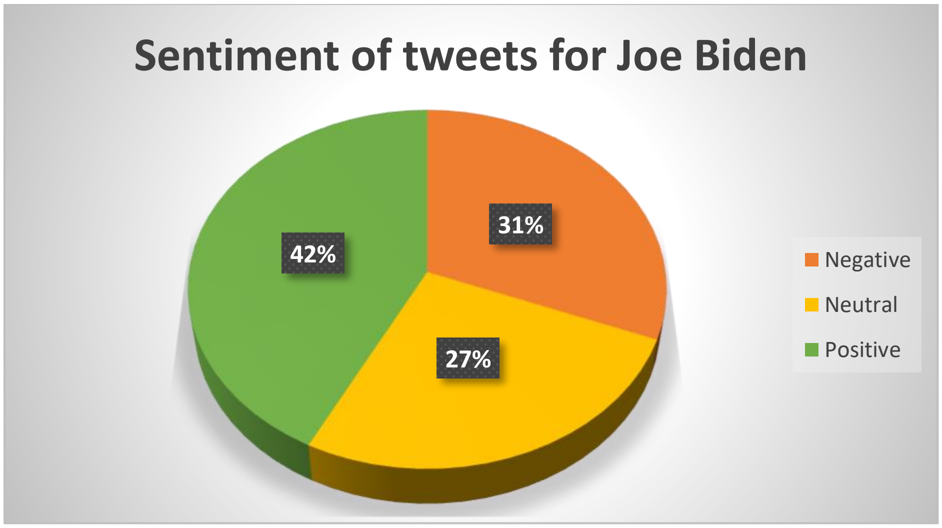


Figure 4: ***Predicted tweet sentiment distribution for Joe Biden***

IV. A further analysis on the predicted sentiment of tweets for Joe Biden shows that the positive tweets were always higher compared to the negative and neutral tweets between 10/25/2020 to 11/03/2020. The number of negative tweets dropped in the last two days from the election date. Figure 5 shows the sentiment trend over the period of the data collection.

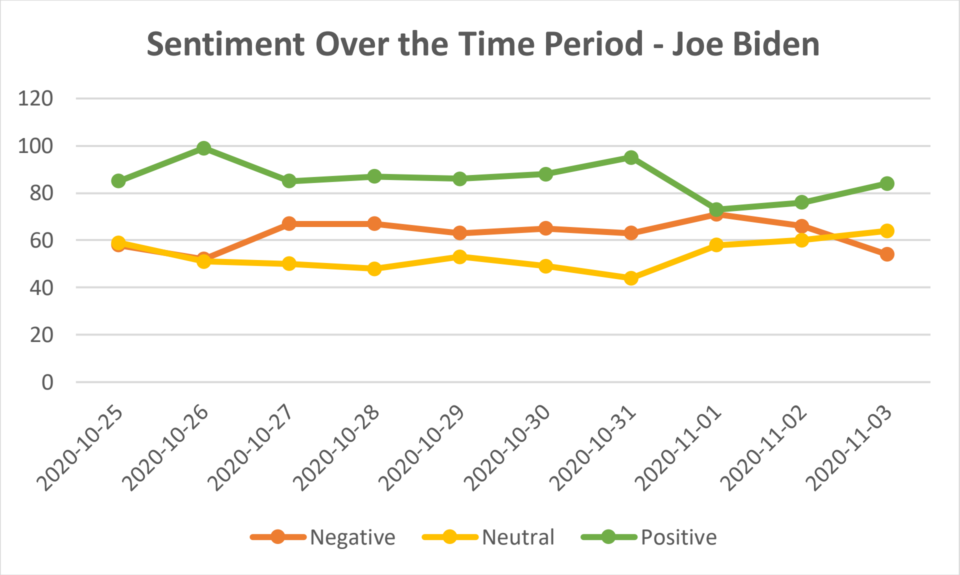


Figure 5: ***Predicted tweet sentiment trend for Joe Biden between 2020-10-25 to 2020-11-03***

(*Trend is based on the combined sentiment counts of Joe Biden and Donald Trump - 2020 data points*)

5. CONCLUSION AND LIMITATIONS

The study intends to find out the sentiment of people regarding Joe Biden running for the President. The Twitter sentiment analysis shows that Biden had a positive impact before the election and most of the people supported with positive tweets for him. Apart from that, the study also shows that people criticized Donald Trump as there were negative tweets most of the time. Joe Biden always led the campaign and people favored him over Donald Trump before the election.

There are obvious limitations to the research, as the amount of data is very limited, and the duration of data collection is short. There is still a scope to analyze the tweets if collected geo-location wise. The model can be further tested by applying it to past or future elections using relevant Twitter data for that span of time, but when further refined, it may be used to predict the results of individual states and the electoral college.

REFERENCES

[1] Shevtsov Alexander, oikonomidou Maria, Antonakaki Despoina, Pratikakis Polyvios, Ioannidis Sotiris. “Analysis of Twitter and YouTube during USelections 2020”. Social and Information Networks (cs.SI). volume 4. 10 November 2020. arXiv:2010.08183v4 [cs.SI] 10 Nov 2020

[2] Guha, Pritam. “Sentiment Analysis on Twitter data regarding 2020 US Elections”. Towards Data Science, 1 November 2020. https://towardsdatascience.com/sentiment-analysis-on-twitter-data-regarding-2020-us-elections-1de4bedbe866

[3] Zhao Jianqiang and Gui Xiaolin. 2017. Comparison research on text preprocessing methods on twitter sentiment analysis. IEEE Access 5 (2017), 2870– 2879.

[4] Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In LREc, Vol. 10. LREC, Valletta, Malta, 1320–1326.

[5] Adam Bermingham and Alan Smeaton. 2011. On using Twitter to monitor political sentiment and predict election results. In Workshop on Sentiment Analysis where AI meets Psychology. 2–10.

[6] Bajaj, Aryan. “Can Python understand human feelings through words? – A brief intro to NLP and VADER Sentiment Analysis”. Data Science Blogathon, Analytics Vidya, 17 June 2021. https://www.analyticsvidhya.com/blog/2021/06/vader-for-sentiment-analysis/