Twitter Data Analysis of 2020 Elections Presidential Candidates

A project report submitted in fulfilment of the requirement for the

BAN 675 – Text Mining

Submitted by

Group 10
Devi Nadimpally (yy4246)
Priyanka Shah (mt8964)
Uttara Dabbiru (py4642)
Venkat Narayanan (fz2722)

Master of Science in Business Analytics

California State University, East Bay

Under the guidance of

Dr. Peng Xie

COLLEGE OF BUSINESS AND ECONOMICS

CALIFORNIA STATE UNIVERSITY, EAST BAY

1. ABSTRACT

The 2020 Presidential election was a landmark political event, contested between the current incumbent Joe Biden of the democratic party and Donald Trump, of the republican party. The prediction of the United States presidential elections attracted a great deal of attention from researchers to try and predict what factors influenced public in supporting a presidential candidate. Most of the predictions made for the 2016 US presidential election were wrong, which increased the spotlight on the various trends that were used to measure the popularity level of the candidates. The main aim of this project is to propose an approach to observe the social engagement of the candidates, specifically on twitter and to predict which tweets help to grab more attention and have a bigger influence. With the help of various text analysis methods, such as Sentiment analysis, Topic modeling, Sophistication measure, Lexicon sentiment analysis, a Negative Binomial Regression model is built to predict the social engagement trends and provide insights regarding the candidates.

I. INTRODUCTION

The 59th USA presidential election was one of the largest political events that happened in 2020. Held in the backdrop of the COVID-19 pandemic and amid various protests raging across the country, considering the various racial discrimination issues, the election was one of the most important events that was widely anticipated and analyzed by various people across the globe. The presidential election was contested between the current incumbent Joe Biden of the democratic party and Donald Trump of the republican party. Donald Trump was the holder of the office of President of the United States at that time. However, he lost to Joe Biden, who managed to wrestle victory in key swing states in route to his path to the White House. Various news agencies and poll trackers had predicted a close race between the two contestants and at various points during the counting process, it was as such. The election saw a record number of ballots cast early and by mail due to the ongoing pandemic. Many more Democrats voted by mail compared to Republicans. As a result of a large number of mail-in ballots, some swing states saw delays in vote counting and reporting; this led to major news outlets delaying their projection of Biden and Harris as the president-elect and vice president-elect until the morning of November 7, three and a half days after the election.

The main aim of this project is to observe, analyze and predict the social influence of the presidential candidates, based on their regular social activity on the popular media platform, Twitter. Twitter is a popular microblogging and social networking service on which users post and interact with messages, otherwise known as tweets. Registered users can post, like and retweet tweets.

The dataset for the project is outsourced directly from the official twitter accounts of both the candidates by using the python package “Tweepy” which gives access to the Twitter API with python. Around 879 Biden’s tweets and 936 Trump’s tweets, were mined from both the candidates accounts between the timeline of September 3, 2020 to November 3, 2020. From the tweets obtained, it was observed that after removing non-English tweets from each candidates’ tweets, a total of 846 Biden’s and 914 Trump’s tweets were considered for the subsequent analysis.

Furthermore, URL links, usernames and emojis were removed using regular expression from each candidates’ tweets and the existing data frame was updated. A couple of features were performed on this raw data and the remaining features have been performed after word tokenization, removing stop words, and finding root form of words using stemming.

Additionally, new predictor variables were derived from the available variables to do a more meaningful analysis. The feature engineering was predominantly done, and new variables were added such as Tweet length, Compound using Sentiment analysis, FRE using Flesch Reading Ease sophistication measure of reading, 13 Topics using Topic modeling and 4 emotions (Strong, Weak, Hostile and Pleasure) using Lexicon sentiment analysis.

Since, the data frame mainly contained count data, a Negative Binomial Regression model was used to perform statistical analysis and predict the number of likes, retweets, and replies.
The project aims to address the following research questions:

- What types of presidential candidate tweets tend to grab more attention?
- What type of tweets can cause a bigger influence / social engagement?

Presidential tweets are a good factor in understanding the public mood and opinion in different US states, towards the candidates. The engagement of the candidates on social platforms, is nowadays seen as a vital campaigning option, made more essential at a time when the COVID-19 pandemic was ravaging through the United States, heavily impacting campaigning methods like public rallies, which required public outreach.

All predictor variables can serve as important factors in determining the popularity / influence level of the tweets put out by the candidates. With the help of predictors, a Negative Binomial Regression model is constructed which is ideal for both count and over-dispersed data.

II. DATA

2.1. DATA COLLECTION

Twitter data was sourced from the official twitter handles of the presidential nominees of the largest election in the COVID-19 era, conducted in 2020, namely incumbent Joe Biden and his republican rival, Donald Trump. The timeline of the tweets that were mined is from September 3, 2020 to November 3, 2020.

Twitter is an excellent place to extract the data for academic projects as Tweets are a representation of natural language on social media. Twitter API provides consumer keys and authentication tokens. The python package 'Tweepy' makes connection with Twitter by passing authentication information.

Additionally, the developer version of 'snscrape' library helped to extract all data which was required for the research questions.

2.2. DATA DESCRIPTION

TABLE 1. DATA DESCRIPTION.

Feature	Description
ID	Integer variable that refers to the twitter unique ID
Date	String variable that refers to the tweet date
Tweets	String variable that refers to the tweets of presidential candidates
Reply count	Integer variable that refers to number of replies for each tweet
Retweet count	Integer variable that refers to number of retweets for each tweet
Like count	Integer variable that refers to number of likes for each tweet

2.3. DATA CLEANING

Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, irrelevant, duplicated, or improperly formatted. This data is usually not necessary or helpful when it comes to analyzing data because it may hinder the process or provide inaccurate results.

Data Cleaning steps followed for both candidate tweets:

- Removed non-English tweets from dataset.
- Removed html links from dataset.
- Removed @username from dataset.
- Got rid of punctuations, special characters and emojis.
- Created list of words using nltk package, word_tokenize.
- Created stop words list and removed unwanted words from the dataset.

- Utilized Porter Stemmer package to get root of word in the dataset.
- Removed rows where tweet length is 0 after removing stopwords.

2.4. SUMMARY STATISTICS

The describe method in Pandas is used to view basic statistical details like mean, standard deviation percentile and min, max value of the numeric data.

The mean is the average of the dataset found by adding all numbers in the dataset and dividing by the number of values in the set, Median is the mid value of the data set and Mode is the number which occurs most often. Percentiles indicate the percentage of scores that fall below a particular value.

	ID	Reply_count	Retweet_count	Like_count
count	8.460000e+02	846.000000	846.000000	8.460000e+02
mean	1.314772e+18	5607.776596	15034.667849	9.768625e+04
std	6.394409e+15	6613.414468	22888.841267	1.522697e+05
min	1.301316e+18	81.000000	619.000000	2.497000e+03
25%	1.309970e+18	1856.750000	4037.000000	2.067800e+04
50%	1.315777e+18	3302.500000	8114.500000	4.255750e+04
75%	1.320392e+18	6327.000000	17658.000000	1.136115e+05
max	1.323410e+18	52117.000000	303586.000000	1.838618e+06

FIGURE 1. STATISTICS SUMMARY OF BIDEN'S TWEETS.

	ID	Reply_count	Retweet_count	Like_count
count	9.140000e+02	914.000000	914.000000	914.000000
mean	1.312314e+18	823.458425	971.310722	4553.881838
std	6.771670e+15	1080.762276	974.569144	4424.462646
min	1.301328e+18	25.000000	153.000000	730.000000
25%	1.306307e+18	320.250000	427.250000	1982.000000
50%	1.310673e+18	532.000000	679.000000	3308.500000
75%	1.318715e+18	934.250000	1102.500000	5142.750000
max	1.323406e+18	14638.000000	9944.000000	43596.000000

FIGURE 2. STATISTICS SUMMARY OF TRUMP'S TWEETS.

Upon observation, the data was seen to be skewed and a method was required to analyze the skewed data. However, since the Negative Binomial model primarily uses count data, alternate methods like OLS regression model were not preferred.

III. METHODOLOGIES

To analyze the influence and social engagement of the tweets, various features must be taken into consideration. The study ensures a comprehensive analysis by modeling prediction using different feature predictors and a predictive algorithm for count data.

The following features were engineered to enable the prediction model to analyze and predict the count of likes, retweets, and comments:

- Sophistication Measure of writing
- Sentiment Analysis
- Length of the tweets
- Hourly Tweets
- Topic Modeling
- Lexicon Sentiment Analysis

3.1. SOPHISTICATION MEASURE

Readability is the ease with which a reader can understand a document or a text. In natural language, the readability of text depends on the complexity of its vocabulary. It mainly focuses on the words we choose, and how we put them into sentences and paragraphs for the readers to understand.

If a piece of writing fails to convey that information, it is likely because it was not written in a clear, straightforward fashion.

Flesch Reading Ease score is one of the readability score determination methods and formula is considered as one of the oldest and most accurate readability formulae.

Equation (1) is:

$$FRE = 206.835 - 1.015*(\text{total words}/\text{total sentences}) - 84.6*(\text{total syllables}/\text{total words})$$

In this project, utilized the Textstat Python package on cleaned raw tweets which calculates statistics from texts to determine their readability, complexity, and school grade level.

A higher score in Flesch's reading ease test indicates a sentence that is easier to read; lower score mark sentences that are more difficult to read.

Fig:01 shows the distribution of Flesch score and frequency of the tweets for both candidate tweets

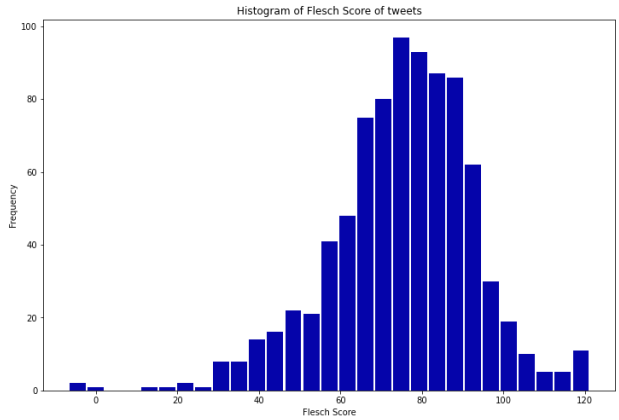

FIGURE 3. FLESCH SCORE AND FREQUENCY OF BIDEN'S TWEETS.

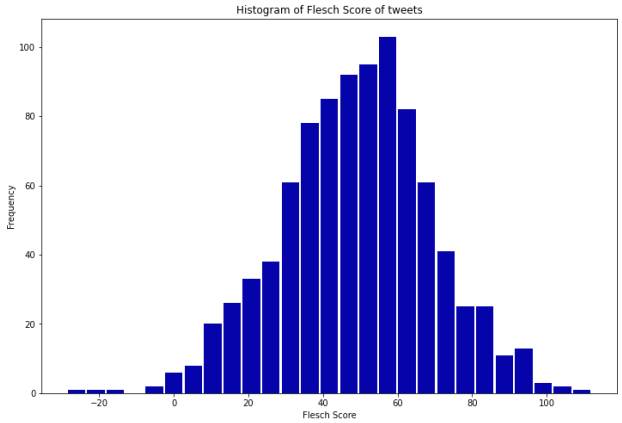

FIGURE 4. FLESCH SCORE AND FREQUENCY OF TRUMP'S TWEETS.

According to 'Fig. 3' and 'Fig 4', Biden's tweets are fairly easy to read while Trump's tweets are fairly difficult to read.

3.2. SENTIMENT ANALYSIS

The sentiment generated by the tweets can help by a defining factor in adjudging the positive or negative reaction from the public towards the candidate's thoughts and views. In this process, cleaned raw text used to calculate positive, negative, neutral, and compound parameters using the Vader Sentiment package. The compound score is varying between the range of -1 to +1. If the score is less than 0, the sentiment is negative. If the score is greater than 0, the sentiment is positive and the score is equal to 0, the sentiment is neutral.

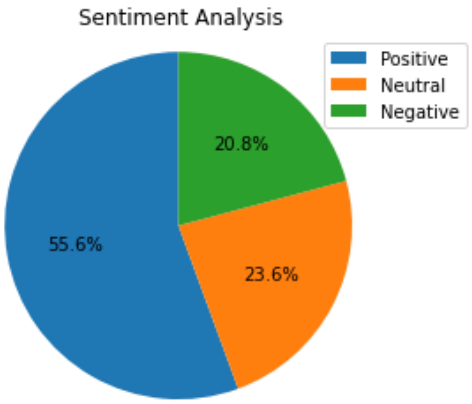

FIGURE 5. SENTIMENT FEATURE RATE OF BIDEN'S TWEETS.

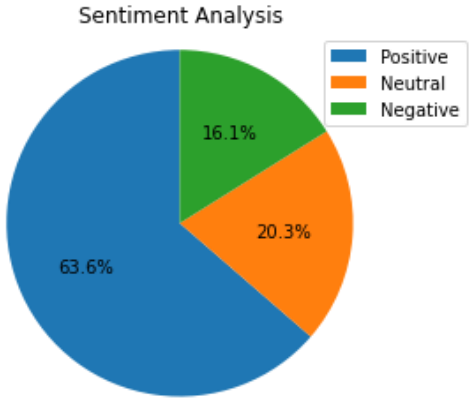

FIGURE 6. SENTIMENT FEATURE RATE OF TRUMP'S TWEETS.

According to 'Fig. 5' and 'Fig 6', 55.6% of Biden's tweets are positive, while Trump's tweets are 63.6% positive in the dataset, which contains more Trump's tweets.

3.3. LENGTH OF TWEETS

The length of the tweets can help as an indicator as to the level of popularity or influence a candidates' tweet has on the public, i.e., the more the length, the less the tweet is popular. The data initially collected, goes through a cleaning process including tokenization, stop word removal and stemming.

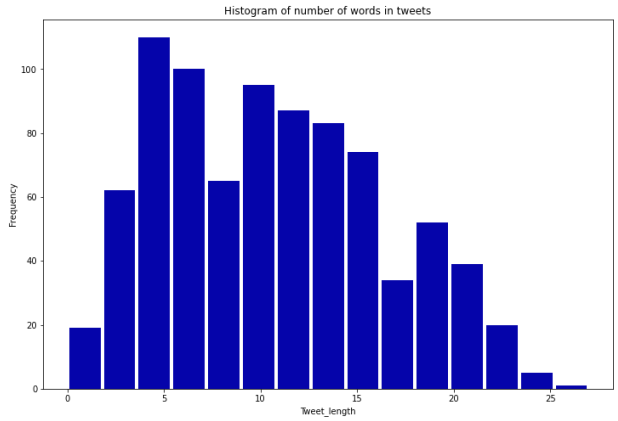

FIGURE 7. TWEET LENGTH AND FREQUENCY OF BIDEN'S TWEETS.

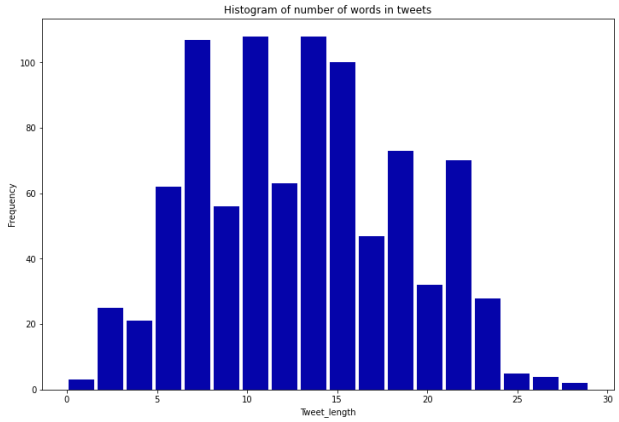

FIGURE 8. TWEET LENGTH AND FREQUENCY OF TRUMP'S TWEETS.

According to 'Fig. 7' and 'Fig 8', a higher number of Biden's tweets have approximately 5 words and Trump's tweets have approximately 10 words.

3.4. HOURLY TWEETS

The hourly tweets feature can help understand when the maximum number of tweets are tweeted. This feature is utilized for explanatory analysis.

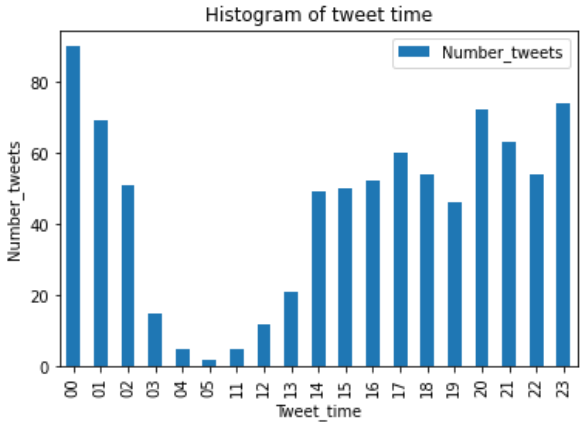

FIGURE 9. HOURLY TWEET AND FREQUENCY OF BIDEN'S TWEETS.

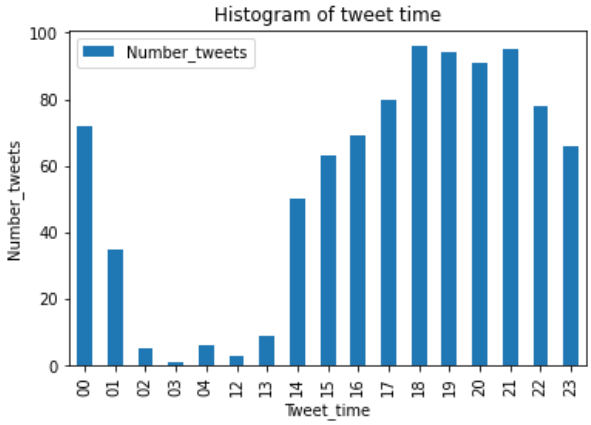

FIGURE 10. HOURLY TWEET AND FREQUENCY OF TRUMP'S TWEETS.

According to 'Fig. 9' and 'Fig 10', most of the tweets are tweeted during the evening until midnight from both the candidates.

3.5. TOPIC MODELING

Topic modeling is the task that discovers topics occurring in a collection of documents. Latent Dirichlet Allocation (LDA) is a way to get the topics of the sentences.

Latent Dirichlet allocation (LDA) from the gensim package is one of the most popular methods of topic modeling. It is used to classify text in a document to a particular topic, it builds a topic per document model and words per topic model. The LDA topic model algorithm requires a document word matrix and a dictionary as the main inputs.

A word cloud was created containing the top 100 words after removing 10 words from tweets which had a frequency more than 100. Using the LDA model, 13 topics were specified, 250 passes (number of

training passes over the dataset) were made and chunk size of 20 is specified for each candidates' tweets. Based on these parameters, topics were distributed to each tweet.

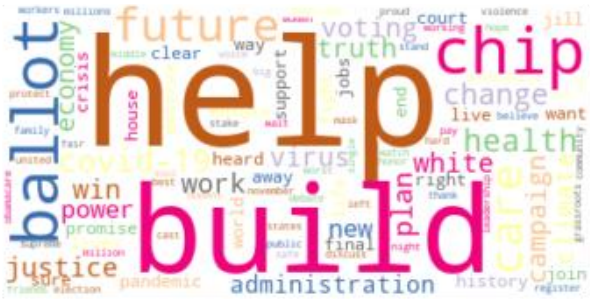

FIGURE 11. WORD CLOUD OF BIDEN'S TWEETS.

FIGURE 12. WORD CLOUD OF TRUMP'S TWEETS.

According to 'Fig. 11' and 'Fig 12', Biden's tweets are more focused on building for the future and Trump's tweets are more focused on his time as the President.

3.6. LEXICON SENTIMENT ANALYSIS

The emotions conveyed by the tweets also helps determine the social engagement the candidates have with the public. In this step, four lexicons were taken containing the emotions Strong, Weak, Hostile and Pleasure. The weightage of each emotion was calculated using sum of words that matched the above lexicons and divided by length of tweets.

IV. REGRESSION ANALYSIS

Negative binomial regression is for modeling count variables, usually for overdispersed count outcome variables. It is one such a model that does not make variance equal to the mean assumption about the data. The dataset contains count of replies, likes, and retweets which are over-dispersed and all explanatory variables as predictors.

To execute the goal of this project, implemented the NB2 model as a part of the Generalized linear model class. It takes the default value of dispersion parameter as 1 but it varies based on datasets. The

auxiliary OLS (Ordinary Least Squares) regression model without constant helped to calculate dispersion parameter value.

The calculated dispersion parameter value was used on the training data set in the NB2 model, and the significant variables were identified based on the p-value.

Below results are interpreted by merging tweets of both the presidential candidates.

TABLE 2. NB MODEL RESULTS.

Features	NB Reply	NB Like	NB Retweet
FRE	0.0226*** (13.891)	0.0301*** (17.097)	0.0290*** (14.884)
Compound	-0.2607*** (-3.539)	-0.1919*** (-2.421)	-0.1623*** (-1.853)
Tweet_length	-0.0317*** (-4.674)	-0.0252*** (-3.395)	-0.0228*** (-2.768)
Strong	1.3058*** (3.694)	0.7190** (1.895)	0.9777*** (2.258)
Weak	5.0626*** (4.247)	4.0727*** (3.223)	4.0053*** (2.746)
Hostile	2.3873*** (1.823)	3.3158*** (2.317)	3.4278*** (2.187)
Pleasure	1.2647 (1.158)	6.0497*** (5.110)	3.3463 (2.546)
Topic 0	-0.6764 (-1.085)	0.8308 (1.207)	0.1408 (0.187)
Topic 1	-0.3891 (-0.511)	-1.0100 (-1.211)	-0.0725 (-0.078)
Topic 2	-0.9999 (-1.266)	-0.2403 (-0.276)	-0.1456 (-0.154)

Topic 3	-0.6764 (-1.085)	0.8308 (1.207)	0.1408 (0.187)
Topic 4	-1.5601*** (-2.032)	-1.5430* (-1.830)	-1.1182 (-1.208)
Topic 5	-0.8205 (-0.997)	-1.2360 (-1.462)	-0.9117 (-0.915)
Topic 6	-1.2798** (-1.711)	-1.5767** (-1.903)	-0.7116 (-0.781)
Topic 7	-0.6764 (-1.085)	0.8308 (1.207)	0.1408 (0.187)
Topic 8	-0.4107 (-0.532)	-0.9825 (-1.148)	-0.2938 (-0.314)
Topic 9	-0.6764 (-1.085)	0.8308 (11.207)	0.1408 (0.187)
Topic 10	-1.6289*** (-2.076)	-1.8454** (-2.164)	-1.3682 (-1.438)
Topic 11	-0.6764 (-1.085)	0.8308 (1.207)	0.1408 (0.187)
Topic 12	-0.6764 (-1.085)	0.8308 (1.207)	0.1408 (0.187)
Topic 13	-0.4138 (-0.449)	0.4192 (0.417)	0.8425 (0.760)
Topic 14	-0.5931 (-0.663)	-0.8035 (0.810)	-0.1122 (-0.103)
Topic 15	-0.0561 (-0.072)	0.4755 (0.553)	0.9600 (1.021)

Topic 16	-2.102* (-2.473)	-2.4983*** (-2.592)	-1.2178 (-1.173)
Topic 17	-1.8779* (-2.415)	-2.5011*** (-2.904)	-1.2680 (-1.332)
Topic 18	-1.5619** (-1.922)	-2.3154*** (-2.620)	-1.8906*** (-1.908)
Topic 19	-0.6764 (-1.085)	0.8308 (1.207)	0.1048 (0.187)
Topic 20	0.3598 (0.349)	2.6606*** (2.515)	1.9305* (1.655)
Topic 21	-0.6764 (-1.085)	0.8308 (1.207)	0.1408 (0.187)
Topic 22	-1.7559*** (-2.166)	-2.1793*** (-2.482)	-1.4321 (-1.468)
Topic 23	-0.6764 (-1.085)	0.8308 (1.207)	0.1408 (0.187)
Topic 24	-2.1802*** (-2.646)	-2.427*** (-2.697)	-1.0404 (-1.040)

Note: ***: $p < 0.05$, **: $p < 0.07$, *: $p < 0.1$

Upon completion of the regression analysis, we can observe that the social influence of the candidates varies depending on different factors and topic of the tweets.

Some observations that can be made based on the results of the analysis of the features include:

- Social engagement of the public is directly proportional to the complexity of the tweets. When the flesch score (which is used as a measure for complexity of a word) for a tweet is high, the tweet has a high number of likes, retweets, and comments.
- Social engagement of the public is inversely proportional to the length of the tweets. When the length of the tweets is high, public reception to the tweets are low, as indicated by low numbers in likes, retweets, and comments.

- The compound variable is used to measure sentiment. It is observed that an increase in the sentiment of the tweets results in a low number of likes, retweets, and comments.
- Certain results were surprising. Strong, weak, and hostile tweets for example, tended to evoke a strong social interaction as the like, retweet and comment numbers were high. This can be attributed to the fact that the public are more responsive to controversial tweets that spur negative sentiments with the public.

Additionally, as part of the topic modeling process, the weight of the topics was calculated, to understand the interests of the public, regarding the various topics related to the candidates that generally trend on social media. The weight of the topics was used in the negative binomial regression model, to predict count for likes, retweets, and comments, when factoring in the topic of the tweets. One of the limitations that was encountered during this process was that only tweets related to COVID-19 and vaccines were significant with regards to the P-value, when predicting the retweet count.

Some observations that can be made based on the results of the analysis of the topics include:

- Inaugural Day Speech: It was observed that the like count was low and reply count was high. This can be attributed to the fact that the inauguration day speech is well scrutinized in both political and public forums and is usually received with mixed reactions.
- Economy, Frontline Jobs and Small businesses: It was observed that social interaction in relation to the topic of economy was low. The United States economy was robust before the pandemic, as unemployment rates were at an all-time low of 11%. Hence, since the beginning of the pandemic, jobs and revenue took a significant hit. This can further be seen in tweets related to frontline jobs and small businesses, which reflected in the poor social engagement the topic had in online forums like twitter.
- COVID-19: The pandemic had also increased the hardships faced by the public, with a huge number of cases and deaths being reported in a daily manner. Hence, the like, retweet, and comment numbers were understandably low.
- Vaccines: At a time when the world was reeling under the COVID-19 surge, news regarding the vaccines came as a breath of fresh air to the public. Thus, social interaction related to the topic of vaccines, were received very well by the public, as evidenced with likes, retweets and comments related to the topic having high numbers.
- Violence: Tweets related to violence are observed to have a negative influence on the public, as likes and comments are low related to that topic. This can be attributed to the fact that events of

racial violence have been a trademark feature during the last 4 years. Hence, tweets related to violence, have not been received well by the public.

- Judge Nomination: The outgoing president had made a controversial move by appointing Amy Coney Barrett as a supreme court judge. This move was not received well by the public, as it was an unprecedented move for an outgoing president, to push through with the nomination for a supreme court judge. Hence, likes, retweets, and comments are low for that topic.

V. CONCLUSION

The prediction of the United States presidential elections attracted a great deal of attention from researchers to try and predict what factors influenced the public in supporting a presidential candidate. Text analysis of the presidential candidates' tweets with the help of various methods, such as number of words in tweets, Sentiment analysis, Topic modeling, Sophistication measure, Lexicon sentiment analysis, a Negative Binomial Regression model was built to predict the count of replies, likes and retweets of both the candidates which defined the social engagement.

In order to grab more public attention, presidential candidates could consider a few options to have more favorable outreach with their tweets such as shorter length of tweets, easy to read language i.e., tweets having fewer complex words, tweets focusing on current affairs and tweets related to negative and weak emotion topics.