

Data Appendix

I. Dataset 1: “hashtag_donaldtrump.csv”

In the “hashtag_donaldtrump.csv” derived from Kaggle, there were 9 variables remaining after the dataset was parsed through and cleaned. Each row in the dataset represents a tweet towards Donald Trump and the unit of observation is “tweet.”

The columns of the data set, uncleaned:

- created_at
- tweet_id
- tweet
- likes
- retweet_count
- source
- user_id
- user_name
- user_screen_name
- user_description
- user_join_date
- user_followers_count
- user_location
- lat
- long
- city
- country
- state
- state_code
- collected_at

The data set, after cleaning:

- created_at
- tweet
- likes
- retweet_count
- user_name
- user_screen_name
- user_followers_count
- state
- state_code

II. Dataset 2: “hashtag_joebiden.csv”

Next, the dataset titled “hashtag_joebiden.csv” also derived from Kaggle underwent the same cleaning process. Each row of the data file represents a tweet directed towards the presidential candidate, Joe Biden. The unit of observation is “tweet.”

The data set, after cleaning:

- created_at
- tweet
- likes
- retweet_count
- user_name
- user_screen_name
- user_followers_count
- state

III. Variables in the Analyzed Datasets

The “hashtag_donaldtrump.csv” and “hashtag_joebiden.csv” have the same variables:

- created_at - The date and time of tweet creation
- tweet- The full tweet text (This is the text data analyzed for the purpose of this project)
- likes - The number of likes
- retweet_count - The number of retweets
- user_name - The username of tweet creator
- user_screen_name - The screen name of tweet creator
- user_followers_count - The followers count on tweet creator
- state - The state parsed from user_location
- state_code - The state code parsed from user_location

IV. Coding Scheme

Data preprocessing began by importing the raw CSV’s and cleaning the data by dropping unnecessary columns and variables with significant portions of null values. Then, the tweets were filtered to only include those from U.S. states. Preliminary exploratory data analysis was conducted to help gain a better understanding of major patterns and anomalies in the dataset. Then using VADER, sentiment scores (negative, neutral, positive, and compound) were calculated for each tweet with the distributions visualized through various graphs. For text analysis, tweets were further cleaned by removing special characters and stopwords, and then lemmatized. TF-IDF vectorization was then used to identify the most significant words, and graphing the unique high-ranking words for each candidate.

V. Exploratory Data Analysis

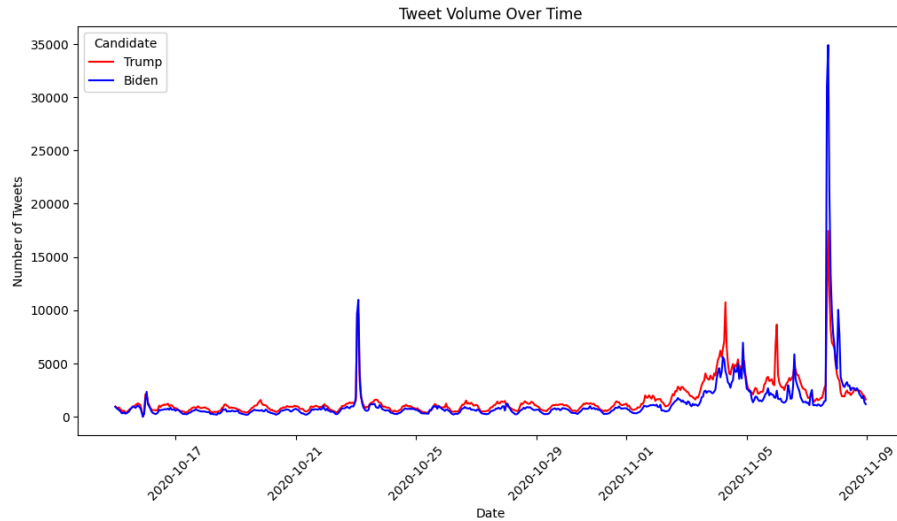


Figure 1. Tweet Volume Over Time

Figure 1 depicts the number of tweets for Biden and Trump in the weeks prior to Election Day. From 10/17/2020 to 11/1/2020, there is not a big difference in the number of tweets towards each candidate. However, after 11/1/2020, there is a small increase in the number of tweets per candidate and the candidates teeter back and forth on who has the most tweets. The graph features a notable peak on 11/08/2020 with a significant number of tweets towards Biden, essentially doubling the number of tweets towards Trump.

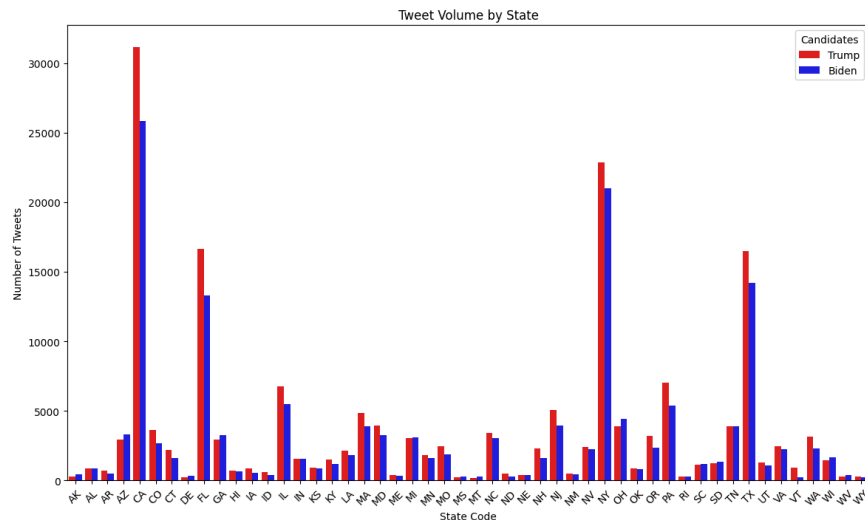


Figure 2. Tweet Volume by State

Figure 2 demonstrates the number of tweets for both candidates again but separated by state. This figure is useful in understanding which states are more involved in the political conversation and voicing their political opinion. Namely, California, Florida, New York, and

Texas stand out and have the highest volume of tweets. It is important to note that these states also happen to be the states with the most electoral votes in the Electoral College.

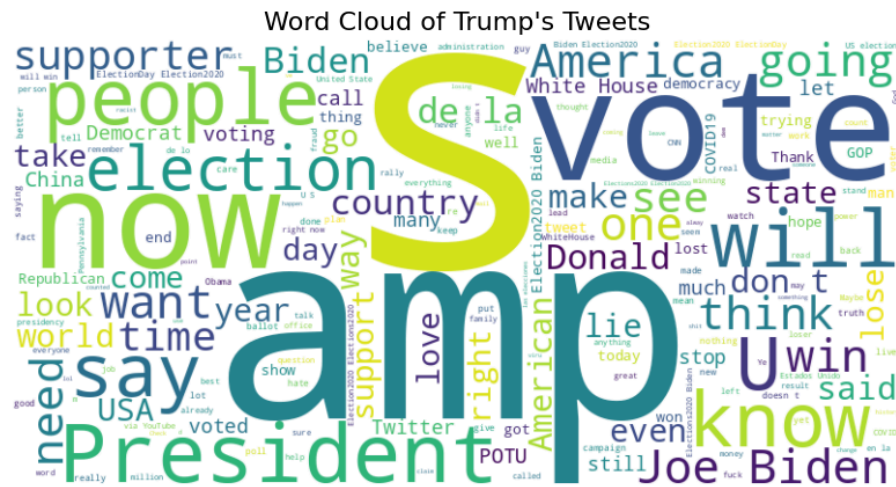


Figure 3. Word Cloud of Trump's Tweets

In figure 3, the words “President”, “amp”, “s” “vote”, “people” and “now” stand out the most. Many of the words on this word cloud are expected and so it is unsurprising why they are populating as the most common words.

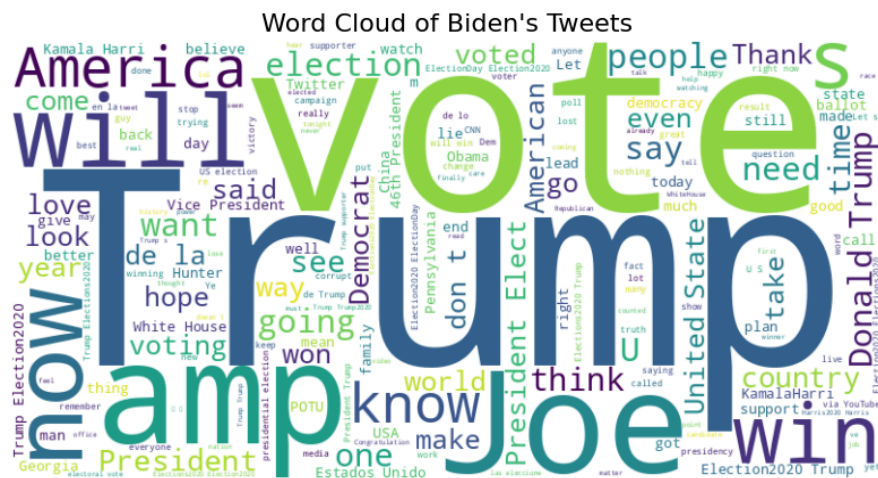


Figure 4. Word Cloud of Biden's Tweets

In figure 4, the words that stand out the most in this word cloud are “Trump”, “vote”, “will”, “amp”, “Joe” and “America.” The word “Trump” jumps out and remains the largest word in this visualization. It is surprising that “Trump” appears most commonly in tweets directed towards Biden.