

September 6, 2020

Sean oneil, susan rico, Quinn E Knudsen

ist 652

Syracuse University

the power of language in politics

uncovering the nuances of political polling and nlp as it relates to the 2020 presidential election

**Table of Contents**

**Section I** ………………………………………………………………………………....pg 1

Project Rationale ………………………………………………………………....pg 1

Section I Roles …………………………………………………………………...pg 1

Data Overview…………………………………………………………………... pg 1

Program Overview………………………………………………………………..pg 1

Investigative Questions Overview………………………………………………..pg 1

**Section I Findings**……………………...………………………………………..pg 2

Question 1………....……………………………………………………...pg 2

Question 2………………………………………………………………...pg 2

Question 3………………………………………………………………...pg 3

Question 4………………………………………………………………...pg 5

**Section II**…………………………………………………………………………………pg 6

Section II Roles ...………………………………………………………………...pg 6

Data Overview…………………………………………………………………... pg 6

Program Overview………………………………………………………………..pg 6

Investigative Questions Overview………………………………………………..pg 8

**Section II Findings**……………….……………………………….……………..pg 8

Question 5………………………………………………………………...pg 8 Question 6………………………………………………………………...pg 11

Question 7……....………………………………………………………...pg 13

Question 8……….………………………………………………………...pg 13

**Section III**..………………………………………………………………………………pg 15

Section II Roles ...………………………………………………………………...pg 15

Data Overview…………………………………………………………………... pg 15

Program Overview………………………………………………………………..pg 15

Investigative Questions Overview………………………………………………..pg 17

**Section III Findings**……….……………………………………………………..pg 17

Question 9………….……………………………………………………...pg 17

Question 10……….……………………………………………………….pg 17

Question 11…………….……………………………………………….....pg 18

Question 12………………………………………………………………..pg 19

**Conclusion**………………………………………………………………………………..pg 21

**Section I**

**Project Rationale**

2020 is a tumultuous year in many regards and a period of radical change in human behavior and society at large. Central to change is policy.  Coincidentally, 2020 is also a year of presidential election in the United States making it a timely endeavor to explore, monitor, and project the manifestation of the pivotal election happening in November. Section II will build upon the analyses conducted within this section by introducing non-structured data derived from Twitter.  Tweets often served as an unbridled, unrestrained modality of communication and may provide insights into how the candidates are faring as well as the sentiment of the nation at large.

**Section I  Roles**

Sean: Report write-up / Preprocessing / Question 1

Susan: Report write-up / Research question design / Question 4

Quinn: Report write up / Research question design / Questions 2 & 3

**Data Overview**

The dataset used for Section I was obtained from the website, [FiveThirtyEight.com](https://data.fivethirtyeight.com/), (presidential\_poll\_averages\_2020.csv) and is made up of 6 columns and over 12,0000 rows. This dataset contains poll averages for the current Democratic and Republican presidential nominees in the 2020 Presidential Election.  The structured dataset did not contain any missing values and only required minimal cleaning. Data was collected from 2/27/2020 to 8/29/2020.

**Program Overview**

We used the numpy, pandas, matplotlib, and plotly libraries. We started by reading in the CSV file with pandas and preprocessing the data. Then we applied different kinds of aggregations to answer our research questions and displayed the results. The plots were made with matplotlib and the tables were shown using plotly.

**Data Preprocessing**

We got rid of the redundant column ‘cycle’ because it only contained a value of 2020. Then we converted the dates from strings into datetime objects so they would display properly. Several “states” called “National” and “NE-2,”, representing regional averages, were removed.

**Investigation Questions Overview**

The following research questions served as guidance as we explored the Twitter dataset:

**Question 1** - How is each candidate trending in the polls over a series of time?

**Question 2** - Which candidate has had a higher estimated voter percentage in 2020?

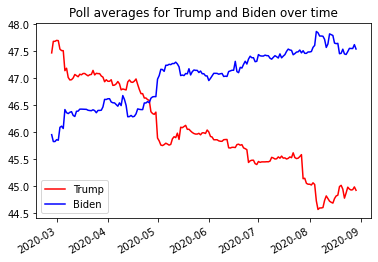
**Question 3** - Which candidate is leading in the key swing states?

**Question 4 -** Which states are a ‘slam dunk’ for a particular candidate to win based on most recent poll data (July 23, 2020)?

**Findings**

**Question 1** - How is each candidate trending in the polls over a series of time?

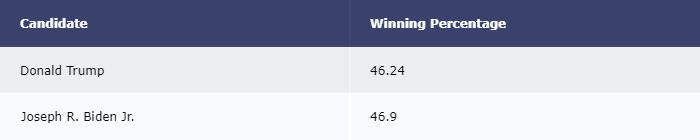
To investigate this question, we grouped each candidate by state and extracted the percent of voters choosing either candidate as a trend in 2020 (Figure 1 below). The results were plotted, and we observed that Trump was initially polling higher than Biden and as time progressed, the candidates shifted in the polls indicating that Trump was trending downward while Biden’s percent of voter estimate began to increase.



*Figure 1.* Trump’s early lead has given way to Biden as 2020 progresses.

**Question 2** - Which candidate has had a higher estimated voter percentage in 2020?

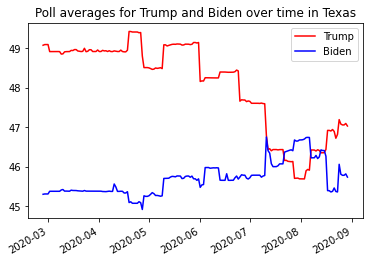
To investigate this question, we grouped by candidate name and calculated the mean percent of votes received for each candidate. Our results indicate that from 2/27/2020 to 7/23/2020, Biden has held an average voter estimation of 46.24%, whereas Trump has held an average rating of 46.9% (Table 1 below).  Although initially leading in 2020, Biden has emerged as the front runner in the more recent months.  Mid-April began the rise of Biden’s estimated winning percentage, with an exacerbated change occurring on April 30th in which Trumps’ odds dropped by 0.48% chance of winning.



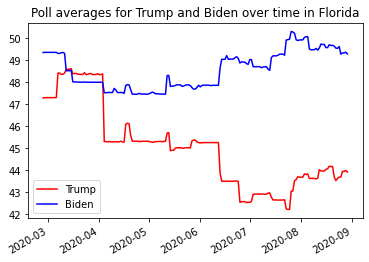
*Table 1.* Over the course of 2020 Biden has been, on average, a head of Trump in the polls.

**Question 3** - Which candidate is leading in the key swing states?

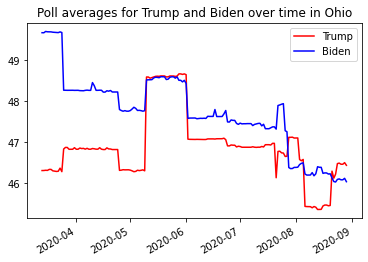
The swing states often decide the election outcome and we wanted to explore some of these pivotal states.  To investigate this highly important information, we took a deeper evaluation of Texas, Florida, Michigan, and Ohio.  The results show favorable results for Biden in all of the states with the closest race being in Texas.



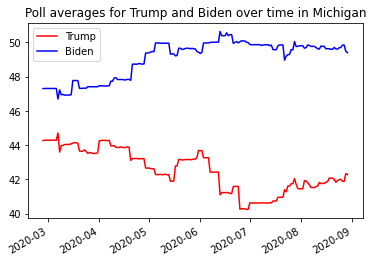
*Figure 2.* Trump’s early lead has rapidly decreased in July before returning in August.



*Figure 3.* Although initially tight, Biden has pulled away in the summer months.



*Figure 4.* Ohio has remained a tight race, although it has favored Biden.



*Figure 5.* Biden has expanded his lead in Michigan as 2020 has progressed.

**Question 4 -** Which states are a ‘slam dunk’ for a particular candidate to win based on most recent poll data (July 23, 2020)?

For this analysis, we calculated the net difference between the percent chance to win for the candidates and then sorted them by the remainder. The majority of these ‘slam dunks’ favor Biden (MA, CA, MD, NY, WA, CT, NJ) while 3 favor Trump (WV, OK, KY).



*Table 2.* This table reflects the states with the highest percent difference between candidates.

**Section II**

**Section II Roles**

Sean: Preprocessing / Analysis / Visualizations

Susan: Data retrieval / Preprocessing/Report write-up/ Analysis

Quinn: Report write up / Research question design / Text analytics / Visualizations

**Data Overview**

The dataset used for Homework 2 was obtained from scraping data from Twitter and is made up of 8 columns (3 computed) and 1737 rows. This dataset contains real tweets from users using popular hashtags related to the Democratic and Republican presidential nominees in the 2020 Presidential Election.  The unstructured dataset required moderate cleaning and removal of columns with many missing values. Data was collected from August 2019 up until August 2020.  Multiple data sets were generated and analyzed based upon whether the goal was to analyze tweets mentioning a certain party of candidate or if the tweet was coming from the candidate themselves.

Reading in the data proved to be a challenging endeavor and several modified approaches were taken.  Initially, we expected to be able to pull in data at various points in time, but the restrictions of the API limited this approach.  Additionally, the sheer volume of tweets containing the hashtags or keywords limited the breadth of time in which we could pull data from. As a result, most of the tweets were a snapshot in time as opposed to a historical view.  Another data issue we dealt with was reading in partial tweets (140 characters) instead of the full length. This was corrected by some additional code specifying full text. Finally, ‘spam’ tweets and quoted retweets slipped through the retweet = false parameter we defined, lowering our sample size and further limiting the scope of tweets we were able to analyze. Spam tweets would be hard to identify, but retweets were removed. This led to an excess in the number of tweets in the final dataset for Republicans compared to Democrats.

**Program Overview**

We used the numpy, pandas, matplotlib, wordcloud, textblob, tweepy, and plotly libraries. We started by reading in the twitter data with tweepy and preprocessing the data. Then we applied different kinds of aggregations to answer our research questions and displayed the results. The plots were made with matplotlib and the tables were shown using plotly.  Sentiment analysis was conducted using the textblob package, giving a numerical value for sentiment for each tweet along with overall categories of positive, negative or neutral. The word cloud was built after removing English stop words and a few custom words such as "https", "co", "amp" as well as some politically tied words that do not add insights to the analysis such as the candidate names.

1.     **Data Retrieval-** We began the process by using the Tweepy API to collect tweets matching a set of chosen hashtags over a recent time period which was saved in a JSON format. We used popular Republican (#Trump, #MAGA2020, #Republicans) and Democrat (#Biden, #BidenHarris, #Democrats) hashtags and put these categories into the final data frame for each tweet. In total, we started with 6000 tweets.

2.     **Data Filtering-** The data was scrubbed in a variety of manners to prepare it for analysis. The JSON data was loaded, and a filter function was designed to remove quoted retweets, any tweets missing the hashtags we searched for, and any tweets with hashtags from both parties which would be ambiguous for the sentiment analysis. After applying this filter, we were left with 1737 tweets.

3.     **Remove Extraneous Variables-** We summarized the user location information in the tweets, originally hoping to only include data from users in the US, but the majority of users left their location information empty, so we decided to exclude this variable. There were many other pieces of metadata included in the JSON, but the variables we kept were the date and time, full text, number of favorites, number of retweets, and number of followers by the user who posted the tweet, along with the category of “Republican” or “Democrat” based on the key hashtag included. This information was stored in a data frame for analysis.

4.     **Execute Sentiment Analysis-** Links and special characters were removed from the tweet text by applying a regular expression, and then sentiment analysis was performed on the text for each tweet using the textblob package. Numerical sentiment scores as well as categories of positive, neutral, and negative were calculated for each tweet and added to the data frame.

5.     **Text Processing-** For the word cloud, we applied tokenization, punctuation and stop word removal, and lemmatization to the text data and generated a frequency table for the different categories.  Custom words were also removed so that things like URLs would not show up in the word cloud.

6.     **Summarize, Bin, and Feature Engineer-** The average sentiment score for each political party was calculated and the two groups were compared with a two-sample t-test. Aggregate counts of each sentiment category were also found for the two parties.

7.     **Visualize Key Findings-** Histograms of the tweet sentiment for the two parties were created, but they looked quite similar, showing many neutral tweets and outliers with large positive or negative sentiment. A bar chart was created showing the count of tweets with negative, neutral, and positive sentiment for the two parties for comparison. A word cloud of the preprocessed tweet text from both parties was created.

**Investigation Questions Overview**

The following research questions served guidance as we explored the Twitter dataset.

**Question 5 -** What are the most common words associated with the hashtag?

**Question 6** - What is the overall sentiment for each hashtag/political party?

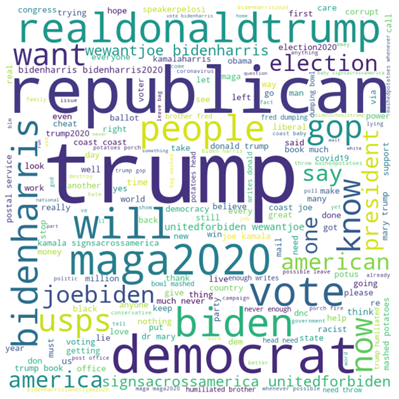
**Question 7** - Do tweets from Donald Trump get more retweets and favorites than Joe Biden?

**Question 8 -** How have retweets and favorite counts trend over time for the two candidates?

**Findings**

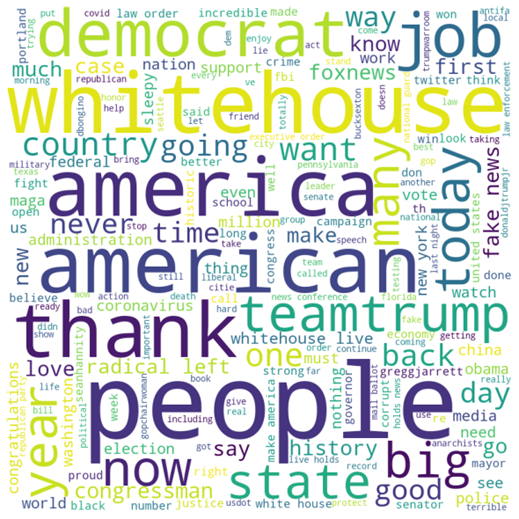
**Question 5 -** What are the most common words associated with the hashtag?

In order to find the most common words used with the presidential candidates we generated a word cloud.  Based upon volume “Trump” and “republican” were used the most often.  As an Easter egg we also discovered “mashed potato” was used enough to appear in the word cloud.  From this word cloud, it is apparent that Twitter users are tweeting the most about Donald Trump and republicans.  The volume might indicate an edge for his campaign, but we will soon explore that further by examining the sentiment around the words being said to discern volume and approval.

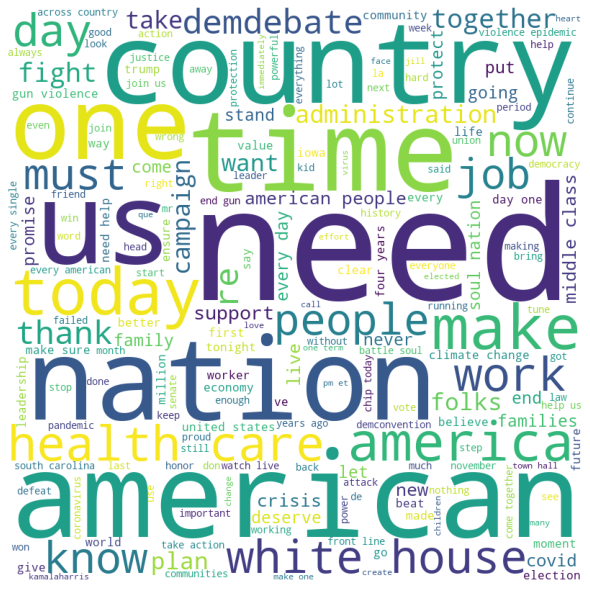


*Figure 1.* A word cloud of the most common words associated with tweets referencing key words associated with the republican and democrat campaigns.

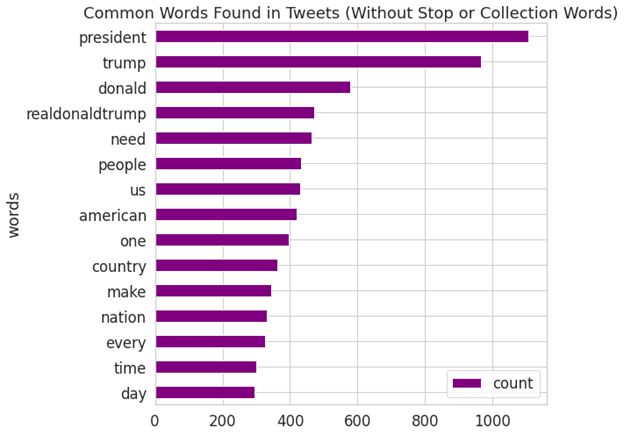
For the individual candidate tweets, additional stop words were removed to avoid the candidate names (Joe, Biden, Trump, Biden). After looking from an outward in perspective, we also explored looking at the words most associated with the tweets coming from the candidates. The last 3,000 tweets from each candidate were pulled and visualized the word clouds below.  In figure 2, you can see that Trump’s  tweets focus on the Whitehouse, thanking the people, jobs, military, campaign slogans, the democratic party, the news (fake news and various channels), coronavirus, and words related to the protests such as law and order, Portland and police. On the other hand, Biden's tweets talk about health care, jobs (the economy and the middle class), the ongoing debates, climate change campaign slogans, COVID, gun laws, and softer future oriented language such as (support, together, promise).  From exploring the volume of tweets through these word clouds, you are actually able to get a good sense of the candidate priorities and where they differ on focus (Trump on law and order, military, the economy and the illegitimacy of the news platforms and Biden on health care, the environment, gun laws, and becoming elected).



*Figure 6.* A word cloud of the most common words associated with Trump tweets from May 21st, 2020 to August 20th, 2020.



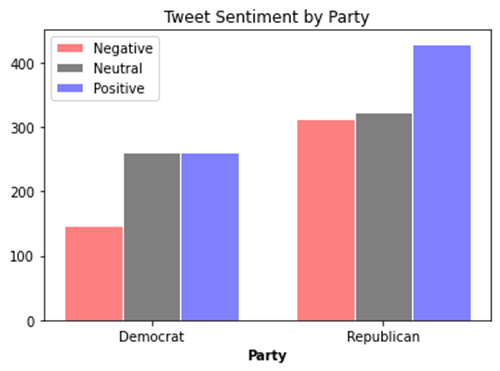
*Figure 7.* A word cloud of the most common words associated with Biden tweets from August 28, 2019 to August 20th, 2020.



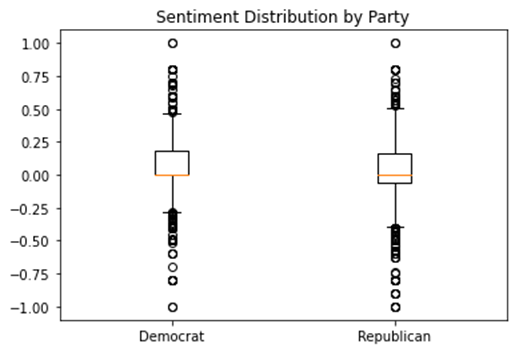
*Figure 8.* A bar chart of the most common words associated with candidate tweets from August 28, 2019 to August 20th, 2020.

**Question 6 -** What is the overall sentiment for each hashtag/political party? Is there a statistical difference between the parties?

Our next question expands upon the volume of tweets by deciphering the meaning behind the words being typed.  A sentiment analysis was conducted scoring overall tweets as either positive, negative, or neutral. As a whole, the tweets were surprisingly more optimistic and positive than negative for both parties.  The aggregate sentiment scores created were then compared to one another through a t-test. Results indicated that tweets that mention democrats had a higher sentiment overall of .06 as compared to the .03 for republicans (p<.05). As displayed below, republicans had higher positivity, but also higher neutral and negative tweets relatively speaking.  Democrats had less tweets overall, but relatively less negative tweets which was the deciding factor for the overall sentiment score. This indicates that republicans are talked about more often and are also more of a polarizing topic than democrats.  When people talk about democrats, there is less negativity shared which may indicate a trend prior to the election and should continuously be monitored.



*Figure 9.* A bar chart displaying tweet sentiment by party



*Figure 10.* A box plot displaying sentiment distribution by party

**Question 7 -** Do tweets from Donald Trump get more retweets and favorites than Joe Biden?

Average favorites for Trump: 96,453.07

Average favorites for Biden: 33,269.89

Average retweets for Trump: 22,296.34

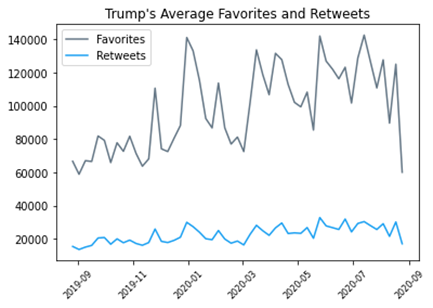
Average retweets for Biden: 6,546.78

To answer this question, we extracted 3000 tweets, which was the max data we could obtain using the Twitter API for each candidate. While exploring our candidate tweets, we also observed a difference in date ranges of the tweets scraped for each candidate.  Although the parameter for tweet extraction was limited to approximately 3000 tweets for each candidate, we observed that in those 3000 tweets, the oldest tweet from Trump’s account was on May 21, 2020 and Biden’s oldest tweet was dated August 28, 2019.  This indicates that Donald Trump is tweeting more frequently than Joe Biden. To ensure that we had the same date range of tweets for both candidates we were able to extract additional Trump tweets in a JSON file from the website [Trump Twitter Archive](http://www.trumptwitterarchive.com/archive), Based on our results, we observed that Donald Trump has averaged more “retweets” and “favorited” tweets in comparison to Joe Biden.  Donald Trump’s average retweets were 22,296 (vs 6,547 for Biden) and his average “favorited” tweets were 96,453 (vs 33,269 for Biden.  However, it is important to note that Biden was unofficially declared as the candidate for the Democratic Party in March 2020, with the official acceptance as Democratic candidate taking place on August 19, 2020. Meanwhile, President Trump has been in office since 2017, possibly indicating a higher level of interest and exposure.

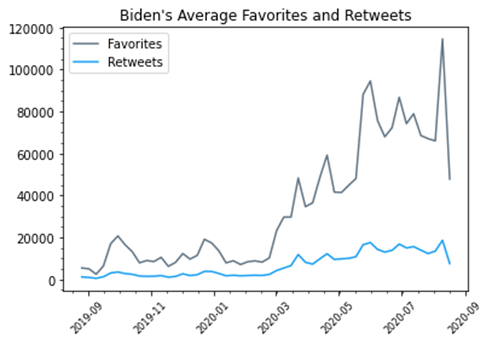
**Question 8 - -**How have retweets and favorite counts trend over time for the two candidates?

Although it was apparent that Trump’s tweets were generally more popular than Biden’s, we wanted to see if there were any trends over time for each candidate in regard to retweets and favorited tweets. We plotted the average number of “retweets” and “favorites” by week for each candidate to see how these were trending. President Trump’s figure displays some sharp spikes and drops in favorited tweets beginning in January through mid-August.  His retweets range between 10k and 30k throughout the 52 weeks of data. Our figure for Joe Biden’s average favorites and retweets shows that while he generally receives less favorites and retweets, his numbers have been increasing over time.   March was an important month for the candidate after he won several key states placing him as the most likely candidate for the Democratic party, which appears to coincide with the uptick in average favorites and retweets.

As a note, candidate tweets that were retweets of other user tweets were removed.



*Figure 11.* A time series plot displaying average Twitter favorites and retweets for Trump



*Figure 12.* A time series plot displaying average Twitter favorites and retweets for Biden

**Section III**

**Section III Roles**

Sean: Preprocessing / Analysis / Visualizations

Susan: Data retrieval / Preprocessing/Report write-up/ Analysis

Quinn: Report write up / Analysis / Visualizations / Data engineering

**Data Overview**

The same datasets used in the prior sections were utilized in this final portion.  Different preprocessing techniques and data manipulations such as grouping the tweets by candidate and by the average over a one week span was used.  Also, features were created such as taking the difference between Trump and Biden poll estimations and the difference between Trump and Biden twitter favorites.  These were used for comparison and for the correlation test.  Finally, the datasets were also joined using the created week data time field to compare the trend of net difference in Trump vs Biden poll percent with the difference in favorites over a 26 week timespan.

**Program Overview**

We started by importing the necessary libraries: numpy, pandas, re, nltk, sklearn, matplotlib, plotly, datetime, scipy, and networkx. Then, for each question, we loaded in the data, applied various preprocessing, ran analyses, and produced visualizations. The details of the program implementation are explained in each question, and the code is also generously commented.

1. Generate and tune a predictive model.
2. Create key calculated fields.
3. Visualize a choropleth map.
4. Inner join Twitter and FiveThirtyEight datasets.
5. Test for significant relationships and visualize scatterplot.
6. Discover and visualize candidate bigrams.

**Investigation Questions Overview**

The following research questions served as guidance as we explored a synthesis between the Twitter dataset and FiveThirtyEight poll estimation datasets.

**Question 9  -** Can we predict which candidate wrote the tweet based upon the language?

**Question 10 -** How is each candidate performing in the polls throughout the country?

**Question 11 -**  Is there a relationship between Twitter favorites and polling estimates?

**Question 12  -** Which phrases did the candidates use most often?

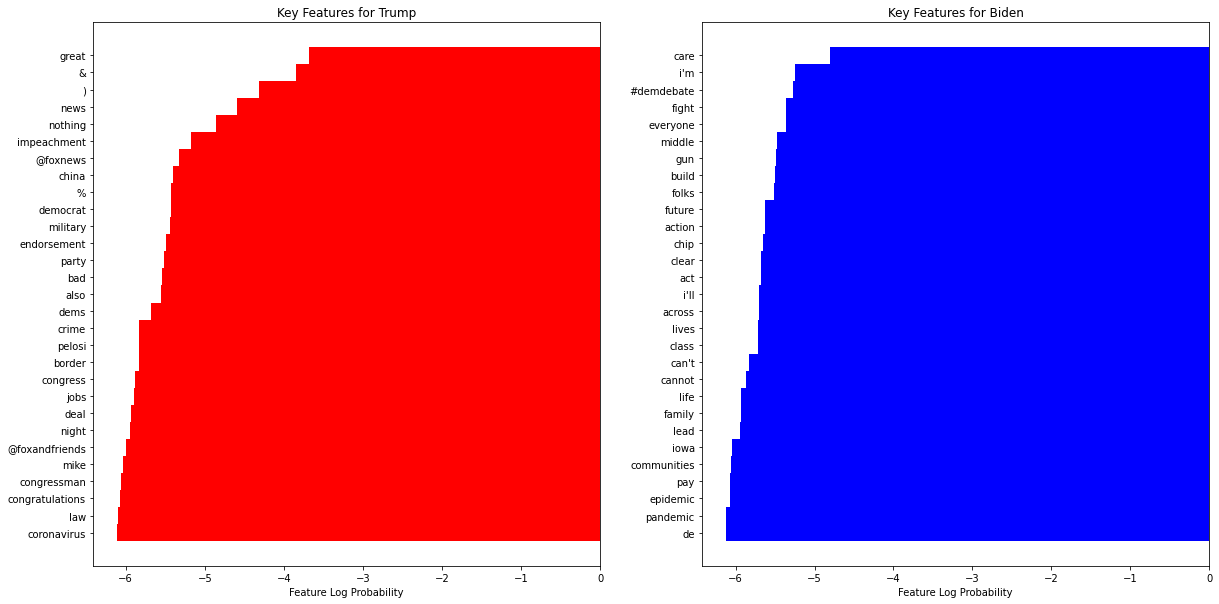
**Findings**

**Question 9  -** Can we predict which candidate wrote the tweet based upon the language?

We felt that since the problem of fake news has become so prominent lately, it may be helpful to be able to predict authorship of a quote allegedly from Donald Trump or Joe Biden. The goal was to build a machine learning classifier which could predict whether a quote was written by Trump or by Biden. We processed the tweet data collected from Donald Trump and Joe Biden starting in August 2019, assigning labels to the candidates arbitrarily (class 0 for Trump, class 1 for Biden). The dataset was slightly unbalanced, with 5518 tweets for Trump and 3000 for Biden. The tweet tokenizer was used, and links were removed using a regular expression. We tried using stemming, but it did not improve the models significantly, so we left it out of the final model. After tokenization, the data was put into a term-document matrix using SciPy’s CountVectorizer. A frequency threshold of 20 was applied to the features to remove infrequent words from the prediction model.

Several different model types were tried, along with different hyperparameters, but the model we settled on was Multinomial Naive Bayes, with a smoothing parameter (alpha) of 0.5. All models were trained and tested with five-fold cross-validation and compared primarily in their precision, but recall was also calculated. The final model had an average precision of 90.7% and an average recall of 90.3%. This is a fairly good model, considering it is only using differences in vocabulary choices (as a bag-of-words model) between the two candidates and not more advanced natural language processing methods.

Next, the top 30 features representative of each candidate were found, excluding features common to both candidates (within the top 50), ranked by their log probability. The plots below show these features. We can see that some of Trump’s most identifying words are “great”, “news”, “china”, “democrat”, “military”, etc. while Biden has words like “care”, “everyone”, “gun”, “folks”, “future”, etc.



*Figure 15.* A comparison key features for Trump and Biden

**Question 10 -** How is each candidate performing in the polls throughout the country?

For this research question, several preprocessing steps were needed to produce a meaningful visual.  First we calculated the percent poll difference (net diff) between Donald Trump and Joe Biden.  The formula for this calculation is:

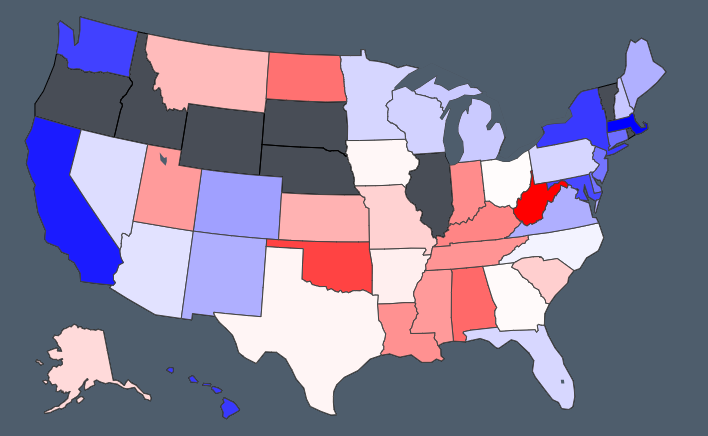
Trump poll percent - Biden poll percent = net difference

In this way, poll differences which favor Trump will have positive values, and those which favor Biden will have negative values. This feature was then utilized in the development of a choropleth map and takes into account third party votes which is meaningful to incorporate.

In its raw form, the state data included the written full text version (e.g. New Jersey) whereas the package we used for mapping the data required abbreviations.  In order to make that work, we used a dictionary of state names and abbreviations to add the abbreviations to the data frame. The resulting field allowed for seamless utilization of the mapping function.

The colors on the map were rescaled so that 0 percent difference was white, while positive values were red and negative values were blue, for Republican and Democrat leads, respectively. States for which data was missing were colored gray.

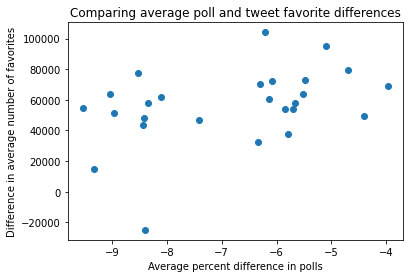
Our figure visualizes the percent difference in polls by candidate, where the redder states tend to favor Trump and the bluer shades favor Biden.  Historically, many states have been consistent with their voting records.  For example, California typically votes Democratic while several southern states vote Republican, which is also reflected in our map visual based on the shade of color.  While this information may not directly provide insights into who will most likely win each state, it does show several states with a smaller percent difference including Texas, Arizona, Nevada, North Carolina and a few others where percent difference is minimal.  This indicates that candidates may want to shift campaign efforts towards these states where polls percentages are close. Some of these states carry high electoral college votes and gaining support may be detrimental for each of the candidate’s chances at winning.

*Figure 13.* A map displaying the predicted 2020 election results

**Question 11 -**  Is there a relationship between Twitter favorites and polling estimates?

For this analysis we had to combine the datasets together. We filtered away data from the tweets so that it matched the same time period as the poll data (Feb. 27, 2020 to present), and this left us with 2697 tweets from Trump and 1530 from Biden. We calculated the difference between Donald Trump’s and Joe Biden’s national poll averages in the same way as the previous question, but aggregated over weeks of the year. The same mathematical preprocessing was done in calculating the differences between the two candidates’ average tweet favorites over time. Aggregation on a weekly rather than daily basis was done in order to help smooth out the variation in average tweet favorites over time. The poll differences and favorite differences were combined together into a new data frame.

This was then tested for statistical correlations and visualized in a scatterplot.  The analysis yielded a small, but statistically significant positive correlation between poll estimates and Twitter favorites.  Based on the R²value from the linear regression, 15% of the variation in candidate poll average differences is accounted for by differences in tweet favorites. This indicates that Twitter favorites could be a potential source of information parties monitor when evaluating the likelihood of their candidate to win and it should be included as a potential feature in any model built for predicting election results.



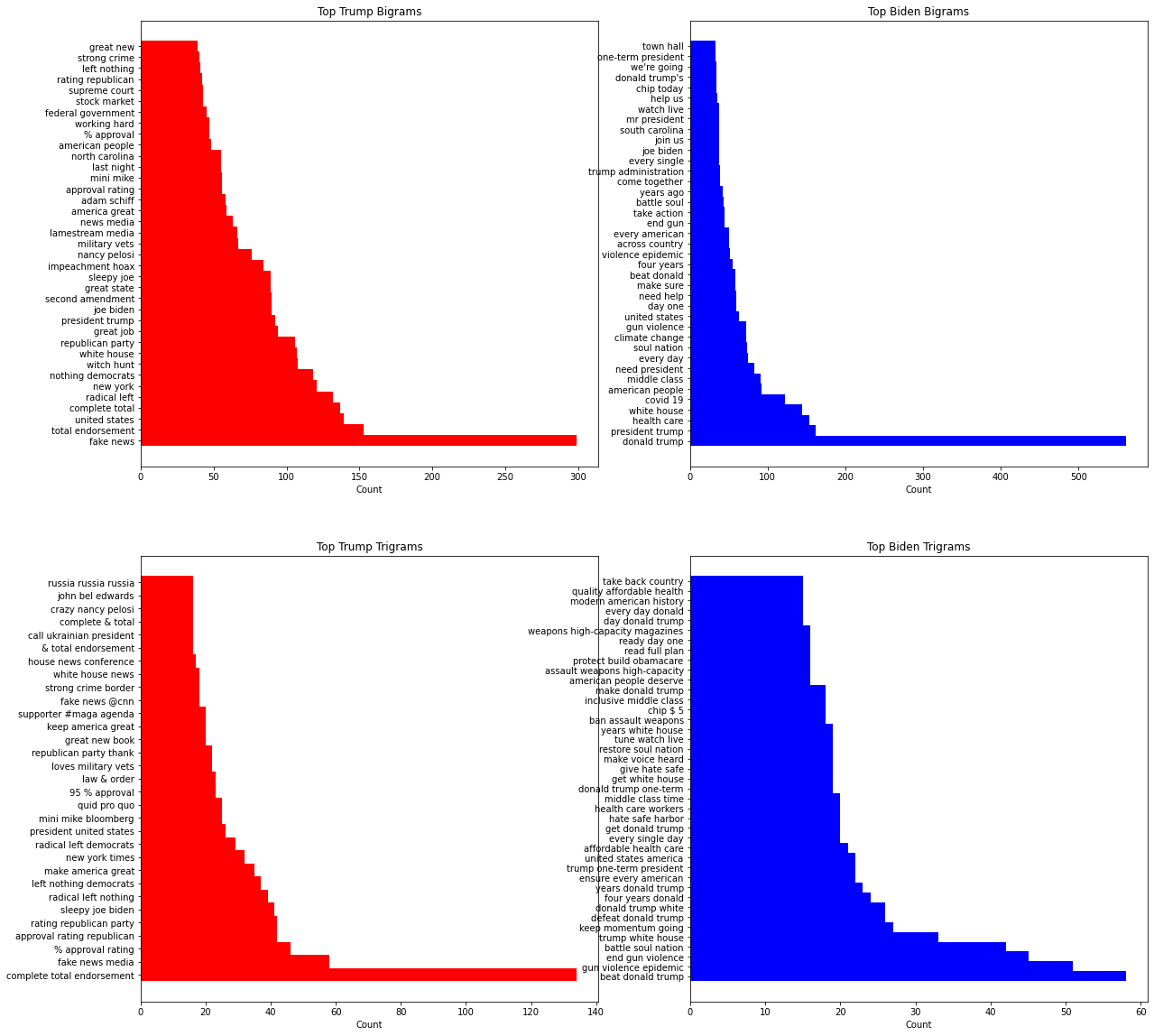
*Figure 14.* Comparison of average polls and favorite difference for each candidate

**Question 12  -** Which phrases did the candidates use most often?

We wanted to find the most common bigrams and trigrams for the two candidates, to see the frequency of common phrases, like “fake news”. To do this, the tweet data for each candidate was tokenized and stopwords and most punctuation was filtered out. NLTK was used to get frequency dictionaries for the bigrams and trigrams for each candidate. The bigrams and trigrams, along with their counts, were collected into lists and then displayed with matplotlib.

From this, we can see that the most common phrases for Trump included things like “fake news”, “complete [and] total endorsement”, “percent approval rating”, “radical left”, etc. while common phrases for Biden were things like “beat Donald Trump”, “gun violence epidemic”, “keep momentum going”, “health care”, etc.

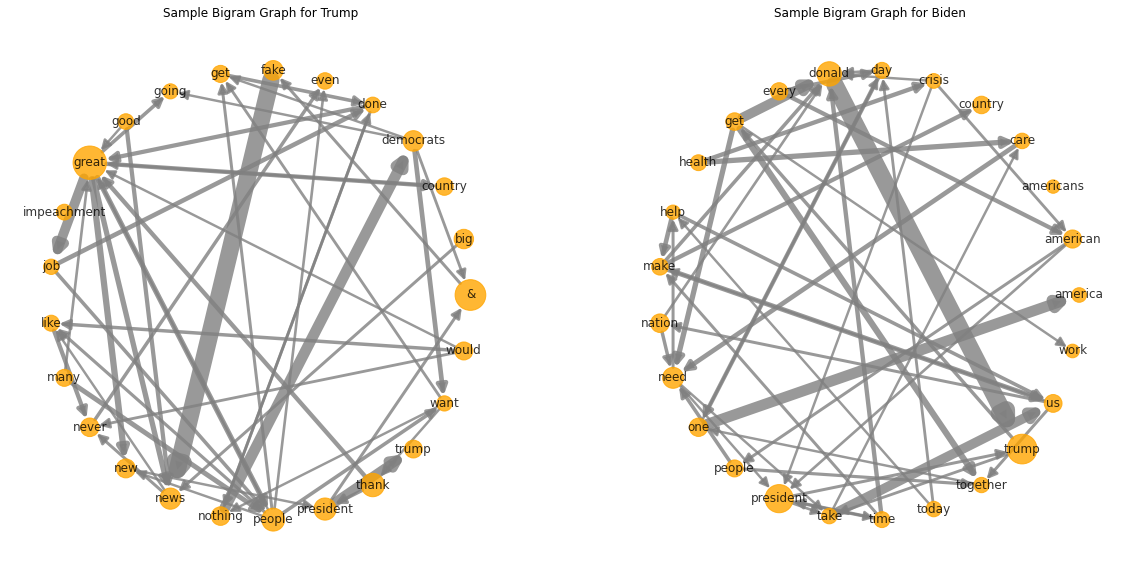
Furthermore, this utilization of bigrams and trigrams demonstrated incremental improvements beyond the single word analyses discussed previously.  The additional context helped explain with greater insights and intricacies into the prioritization areas for the candidates. For example, the trigrams helped discover the frequency at which Trump talks about approval ratings and the monikers he creates for his political adversaries. It also revealed the frequent commentary Biden provides around Trump is language oriented towards removing him from the Whitehouse come November.  The depth and strength of Biden’s stance against the current gun laws was also emphasized through both the variety and volume of ways in which he talks about the subject.



*Figure 16.* Top bigrams and trigrams for the two candidates

Networkx was used to create graphs of some of the common connections between these words. The top 25 words for each candidate (excluding stopwords) were used as the nodes, and the edges were created from the bigram list, with a minimum frequency threshold of 5. The edges were drawn with a thickness depending on the square root of the bigram frequency, and the nodes were sized by word frequency. The graphs were displayed in a circular layout in order to better see the connections better than if they were a single large cluster.

The results (visualized below) reiterated the divergent speech patterns of Joe Biden and Donald Trump.  The top bigram associations for Trump were “Fake News” and “Nothing Democrats” whereas Biden’s top bigrams were “Donald Trump” and “One America”. Another finding was that Trump uses the amplifier “great” with a multitude of words such as “job”, “news”, “country” and “people”.  Interestingly, Biden does not have top bigrams that utilize amplifying words indicating a difference not only in the content that is presented, but also the manner in which each candidate goes about conveying it.



*Figure 17.* Graphs of some of the bigrams for the two candidates

**Conclusion -**

From our initial analysis of the polling data, a clear trend emerged in 2020.  Donald Trump began the year with a sizable campaign lead, but quickly relented to Joe Biden at the beginning of the pandemic.  As the virus spread and national lockdowns persisted, Joe Biden widened his gap in poll estimations with a particularly large change happening on April 30th.  In 2016, Hillary Clinton won the popular election, but was not the presidential nominee due to the electoral college.  We then decided to explore the highly pivotal ‘swing states’ that can either make or break an election.  These states included Texas, Florida, Michigan, and Ohio.  The results initially showed favorable results for Biden in all of the states with the closest race being in Texas. However, over the course of this project, Donald Trump made noticeable gains and even surpassed Biden in a few states.  Finally, we calculated the net difference between the percent chance to win for the candidates and then sorted them by the remainder.  This was calculated to find the ‘slam dunk states’ that are unlikely to swing prior to the election.  This is useful to know because campaigns spend millions of dollars and need to be highly targeted with their funds. The majority of these ‘slam dunks’ favor Biden (MA, CA, MD, NY, WA, CT, NJ) while 3 favor Trump (WV, OK, KY).

Utilizing Twitter data, we found some interesting trends in the data that serve precursors to the election in November and can adequately compliment some of the polling data. Our results here have demonstrated polarization around the Trump campaign with high highs and low lows.  The Biden campaign, although flying under the radar in terms of how people talk about it, has much less negative sentiment even when adjusting for the volume of tweets.  This creates a complex question in that is the lack of volume something to be concerned about, or is the higher sentiment indicative of favorable results come November? As a next step for the final project, we will attempt to equate social media popularity to poll popularity and integrate the data and findings from the first two homework assignments.

More broadly, this analysis was helpful at uncovering the different approaches taken by the two main party candidates as the 2020 election nears. Trump actively uses Twitter and has been in the media spotlight longer than Joe Biden has.  Trump’s ‘popularity’ on Twitter far supersedes Biden’s both in terms of followership and reactions as measured by retweets and favorites. Mirroring this, there was a greater volume of tweets around republican tags than democrats with a polarization of sentiment. Despite the lower volume, the sentiment score of democrat tweets was slightly higher than republicans. Finally, the key topic areas were made apparent in some of the medium to smaller sized words in the word clouds from Trump and Biden. Trump clearly is focused more so on law and order and media sources and Biden is focused on healthcare, gun control and the environment. While this does not indicate an actual stance on a topic area, it is a fairly useful way to get a sense of topic area prioritization.

The trends over time indicate that Trump has stayed flat on social media over the course of 2020. At the same time, Joe Biden has rapidly gained Twitter momentum.  This corresponds with Joe Biden’s assent in the polling data we explored in the first homework assignment.  A statistical correlation will be further explored in the final project. This second source of information is providing evidence that Joe Biden has a strong chance of winning the 2020 election.

Deeper analysis of the Twitter data demonstrated that a highly accurate model can be developed in classifying whether a tweet was written by Donald Trump or Joe Biden.  This 5-fold cross validated model exemplified the unique speech patterns based upon significant features uniquely important to each candidate.

The usefulness of Twitter was demonstrated as it related to election momentum.  This was demonstrated through a few steps. By grouping the data by week, the poll estimate data and Twitter data were joined.  Next, the difference in one candidate’s percent likelihood to win was subtracted from the other candidate’s to get one final score. The same score calculation was made for favorites so that Donald Trump’s weekly average Tweet favorites were subtracted from Joe Biden’s.  A statistical correlation was conducted and yielded a positive correlation (0.39, p <.05) between net favorites and net poll estimation. This indicates that you can gain some evidence of how a candidate is doing in the polls simply by tracking their Twitter popularity.

An exploration of Trump and Biden’s bigrams provided additional insight into both the content and the manner in which the candidates convey information. Biden often talks “at” Donald Trump, calling him by name and paints a picture of “one America”.  Trump talks often about “fake news” and the democrats as a whole as doing “nothing”. Trump has a variety of derogatory monikers for his political foes and emphasizes his popularity in the polls. Biden has taken a strong stance in favor of tighter gun regulations and focuses his tweets about Trump on removing him from office.  Interestingly we found the utilization of amplifying words to be a key distinction between the candidates with Trump using the word “great” with a high frequency.

Holistically, this study demonstrated the effectiveness of using social media data to understand how political candidates are fairing in their campaign. Limitations with the free Twitter API restricted the volume of tweets that were able to be analyzed and future analyses would be best served with an upgraded API.  Although we explored 2 potential sources of information, there are numerous web sources that can additionally be added to this analysis that would likely add value such as other social media platforms (Facebook, Instagram), additional polling sites, news feeds, campaign speech sentiment, party volunteer growth, and campaign donations.  Based upon the publicly available data analyzed throughout this project, it lends itself to a favorable outcome for Joe Biden in 2020.