Alain Duplan
Robin Lovell
Harshul Shukla

Comp Sci 571
May 11, 2021

The 2020 U.S. presidential election was one of the most controversial elections of U.S. history. One of the driving factors behind the controversy was the role of social media platforms in giving voices to many previously unheard communities. These same platforms are often abused and manipulated to drive a narrative. Looking at Twitter, for example, it was plagued with fake tweet allegations, being accused of being biased towards the liberal party, and banning/suspending a multitude of users, one of which being the U.S.

Considering all these factors, we wanted to compare how tweets about the election compared to the actual turnout. Was the proportion of democratic and republican users on the Twitter platform a good representation of the voting population? Did the Twitter users display a biased sentiment towards one party? What effects did spam or fake tweets truly have on the population of uses on the platform? To do so we looked at tweets that were made about the election during the core election period and looked at the similarities and differences between those tweets and the popular vote.

Our initial data was pulled from a Kaggle dataset which included tweets that mentioned either Donald Trump or Joe Biden. The tweets were limited to the last two months of the presidential election period and were separated by which candidate was mentioned, with some overlap. Though the initial data included more features, the variables we chose to include were the tweet ID, tweet text, retweet counts, user join data, user follower count, state code, country, and user ID. In total, there were 958,580 tweets in our Trump dataset and 768,423 tweets in our Biden dataset. We also pulled the electoral college results along with the popular vote to be used in the comparison.

Since our dataset was so large, as we were working with over 1.5 million tweets, we had to perform filtering, as have an efficient runtime and lessen the strain on our computers. Using R, we only selected tweets that had users that lived in the United States and had a valid state code. Since location data was user-provided, we had a large magnitude of users with nonsense or nonexistent locations, and since we are looking at the U.S. election, in particular, it made sense only to look at U.S. users. This removed over 75% of our data from both sets. While in R, we removed tweets that were repeated by the same user, this was our attempt to filter any spam or “botted” tweets that removed under 5% of our data. We also removed any users who have listed join dates before 2006, as the platform was only created in March of that year and we didn’t want to include any users that may have manipulated their accounts. Furthermore, using Google Translate’s Python language detection feature, we were able to select only tweets that were listed in English, further removing 10% of Trump tweets and 15% of Biden tweets. Lastly using R, we formatted our data into a relational database, by first merging both sets of tweets, and removing any repeated tweets. We then separated the set into 2 tables, one of which was for users, containing the user ID as our unique identifier, user join date, user follower count, and state code. Our second table was made for the tweets containing the tweet ID as our unique identifier, tweet text, retweet count, like count, and user ID which was a foreign key that referenced our user table. This leads us to have 161,165 unique tweets and 68,773 unique users in our final tables.

We explored two different methods to derive sentiment from the tweets - ASBA (Aspect Based Sentiment Analysis) and VADER (Valence-Aware Dictionary and Sentiment Reasoner). For ABSA, we used the aspect_based_sentiment_analysis python package, which leverages a BERT-based sentiment analyzer. This allowed us to specify certain entities in the text, such as Biden or Trump, for which sentiment was derived. BERT's use of Transformers allows the model to understand interactions between different words in the same sentence - and therefore their sentiment. While this approach seemed like it would work very well for our task, it did not yield the best results. We believe that this is due to the fact that this model was not pre-trained on social media data and also has no way to derive meaning from the complex, abbreviated, and emoji-heavy text of tweets. Since the main goal of this class and this project was not to develop and train an NLP sentiment analysis model, we moved on to another model that was trained on this kind of data, VADER. This model was able to understand emoticons, slang, and various acronyms that are found in tweets. Furthermore, it was specifically pre-trained on the Twitter dataset, making it a great candidate for this task. Since the data was already separated as tweets with #trump or #biden, we considered the presence of the hashtag to mean that the tweet was about that candidate. We then ran VADER on each individual tweet which returned a polarity score between -1 and 1. If the subject of the tweet was Biden and the score was positive or if the subject was Trump and the score was negative, we considered this as a Pro-Biden tweet. Conversely, if the subject of the tweet was Trump and the score was positive or if the subject was Biden and the score was negative, we considered this as a Pro-Trump tweet. Since all of these tweets were collected in the 2 months leading up to the election, we used the simplifying assumption that a negative sentiment for one candidate implied a positive sentiment for the other and that there were no 3rd party candidates being considered

Our visualization includes nine interactive US state choropleth maps with a diverging red-blue color scale. There is one map for each of the following variables: tweet count, like count, retweet count, tweet sentiment count, popular vote count, tweet-vote difference, like-vote difference, retweet-vote difference, and sentiment-vote difference. A state's color represents the amount by which a candidate had a higher value for a given variable: blue for Biden and red for Trump. The "difference" maps represent the amount by which Twitter data differed from the popular vote in favor of a candidate. For example, in the tweet-vote difference map, a dark red state means that there was a much greater percentage of #Trump tweets than Trump votes in that state.

Radio buttons allow users to navigate to any of the maps. The first three maps feature a static pie chart that shows country-wide totals for a given variable. The first four maps feature an interactive scatter plot that shows the correlation between a given variable and the popular vote. Hovering over states and scatter plot points highlights the corresponding state and point, displays a tooltip with state-specific information, and displays four pie charts below the map representing state-specific tweet counts, like counts, retweet counts, and popular vote counts.

All visualizations were made using D3. Our code is stored in a GitHub repository and published using GitHub Pages. Our repository consists of the folders "/Datasets" and "/docs". "/Datasets" contains six CSV files. "/docs" contains nine HTML files -- one for each map.

Within each HTML file, some data is declared as a JSON object the file and some data is asynchronously fetched from a CSV file and stored in a JavaScript array or object. US maps are made using the D3 topojson library and a US states shapefile fetched from d3js.org. Each state's "id" attribute is set to equal its FIPS code. This id is used to reference the state when applying color scales and event listeners. A red-blue diverging color scale is applied by setting the fill of each state to the function d3.interpolateRdBu. This function takes a number between 0 and 1 and outputs a color. MouseOver and MouseOut event listeners are applied to states and scatter plot points, enabling tooltips and state-specific pie charts to be displayed. The map legends use the D3 color, scale, and scale chromatic libraries. The map legends are created with a "legend" function provided for public use by d3 developer Mike Bostock.

From our visualization, we learned there were more tweets about Trump than Biden. This is likely due to Trump's attention-seeking personality and the controversial nature of his presidency. However, #Biden tweets were liked and retweeted more than #Trump tweets and the sentiment overwhelmingly favored Biden. This was true even in states where the popular vote favored Trump, leading us to believe that Twitter's user base has a left-leaning bias. Of all the variables, tweets and sentiment had the highest correlation with the popular vote. Clusters present in the tweet-vote and sentiment-vote scatter plots indicate there may be a weak correlation between tweets and popular vote and tweet sentiments and popular vote.