2020 Election Sentiment Analysis and State Results Prediction

Alex Hamedaninia

MS Data Science, Bellevue University

DSC 680 Applied Data Science

Professor Iranitalab

April 7, 2024

Research Question

Can we predict the results of the 2020 election based on public sentiment analysis on social media platforms?

In this project, I will conduct sentiment analysis on social media data using natural language processing related to the 2020 election to assess public sentiment. I aim to develop predictive models to forecast state-level voting outcomes based on the observed sentiment trends.

Background

Every four years, as the United States gets ready for the next presidential election, the nation becomes a battleground of competing ideas, aspirations, and visions for the future. With candidates traversing the country, each vying for the favor of the electorate, understanding the prevailing sentiments surrounding their campaigns becomes paramount. What does the public think of them? Is it an overall positive or negative sentiment? Can we gain insights into state-specific sentiments towards presidential contenders? In this project, we will attempt to explore this further by conducting sentiment analysis using natural language processing through the vast expanse of opinions expressed through one of social media's top platforms during the 2020 election year: Twitter.

Social media platforms have emerged as powerful tools for facilitating communication, information dissemination, and interaction on a global scale. With millions of users actively engaging in online conversations daily, social media platforms provide a rich source of data that offers valuable insights into various aspects of human behavior, including public opinion, sentiment, and societal trends. Among these platforms, Twitter stands out as a prominent platform for real-time conversations and discussions on diverse topics. The dynamic nature of

Twitter allows users to express their opinions, share news and information, and engage in discussions with others. This continuous stream of user-generated content provides researchers, analysts, and policymakers with a unique opportunity to gain insights into public opinion in near real-time, enabling them to track emerging trends, identify key influencers, and assess the impact of various factors on public discourse.

Data Explanation

For this analysis, I will be using a dataset obtained from Kaggle. This file contains two datasets; one with Tweets containing any relation to Joe Biden, and another for tweets relating to Donald Trump. These tweets were obtained from 10/15/2020 until 11/8/2020, gathered through the Twitter API. It contains the tweet, the date, user ID, their location, the number of likes and retweets, and other various related information.

To prepare the data for analysis, I removed any unnecessary columns to the study, ensured data type regularity, and dealt with missing values. To maintain data privacy, I removed the *user_name* and *user_screen_name* columns, keeping only *user_id* to distinguish the different users. To reduce computation time and to keep only the variables related to the study, I also removed *user_description* and *user_join_date.* As this study has to do with US votership, I also removed any tweets that did not come from within the United States, which included regulating the user location by converting any location depicted as ‘United States of America’ to just ‘United States’. This was the only variant of the location ‘United States’ that was found within the data.

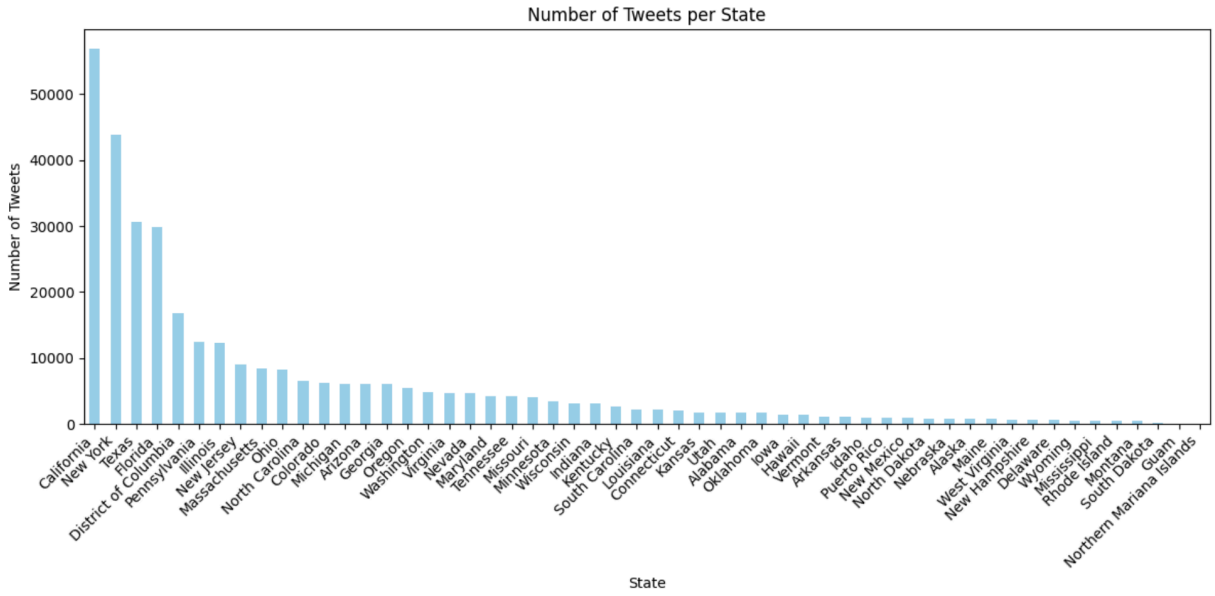
Finally, in dealing with missing values, because we are dealing with a location-based analysis, I created a new subset of data that contained only the rows where the *user_location*

value was not missing. This allows me to conduct the location based analysis without error. After this final step, I created an identifying variable named *candidate* to indicate for each tweet which dataset it came from, and combined the two datasets into one for the analysis portion.

Next steps will involve exploratory data analysis through various plots, followed by the sentiment analysis using TextBlob and finally using those results to predict the winner of the 2020 Election.

Methods

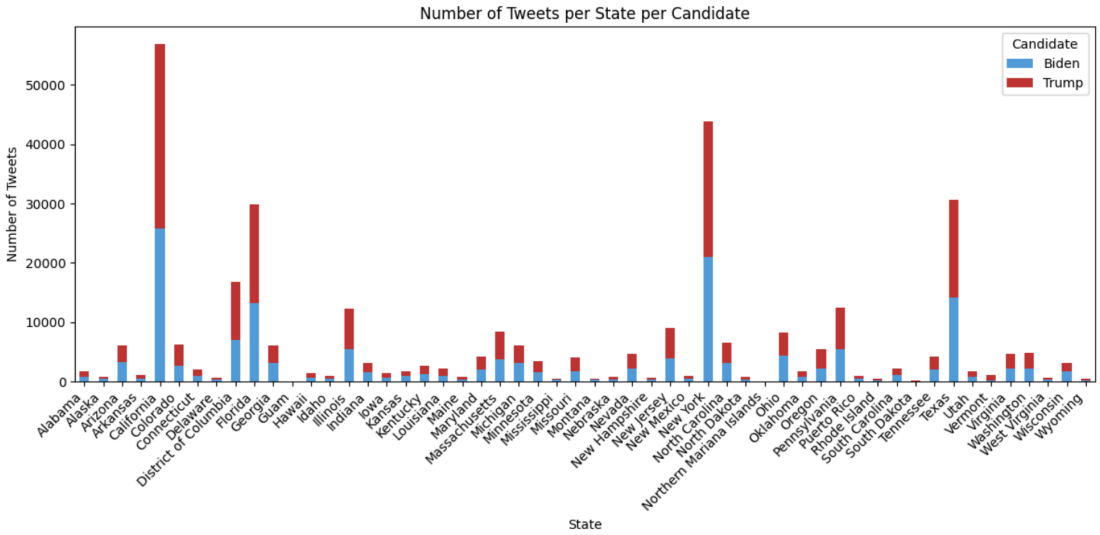
Beginning with the exploratory data analysis, we can gain a deeper understanding of the data. First we can see a distribution of how many tweets come from each state.



We can see there is a very disproportionate amount of tweets coming from the bigger states, due to most likely a higher population. More specifically, these bigger states have a higher city and urban population as opposed to the more rural states. It is important to keep this in mind when conducting further analysis, as a higher tweet amount represents a larger sample size of that

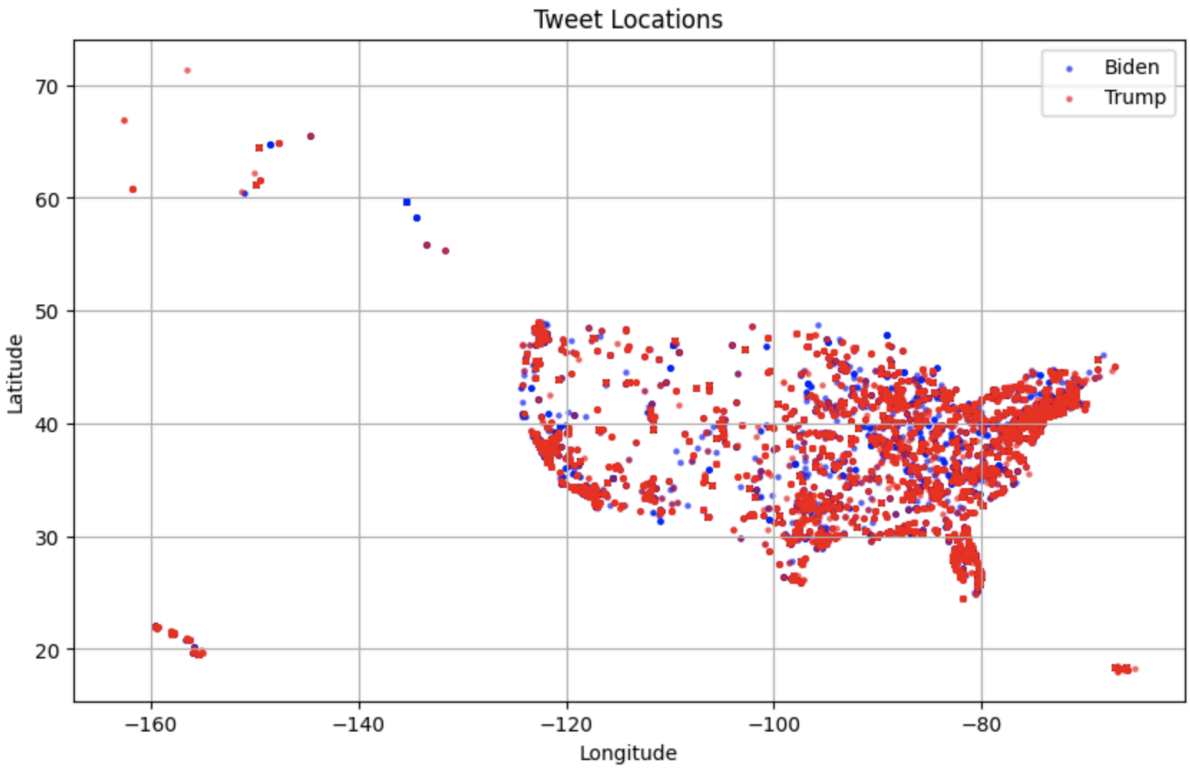
state's population, whereas a lower tweet amount represents a smaller sample size, leading to results that may not fully represent that population.

Next, we can take a look at a similar graph, that shows the amount of tweets regarding Joe Biden versus Donald Trump.



Here, we can see that for almost every state, the amount of discussion surrounding each candidate in each state is nearly equal. While this does not indicate whether the discussion was positive or negative, it does give a clear indicator that for each state, the conversation surrounding both candidates is quite equal.

Next, we can take a look at where the tweets originated from on a map, and see where they are speaking about Biden and Trump.

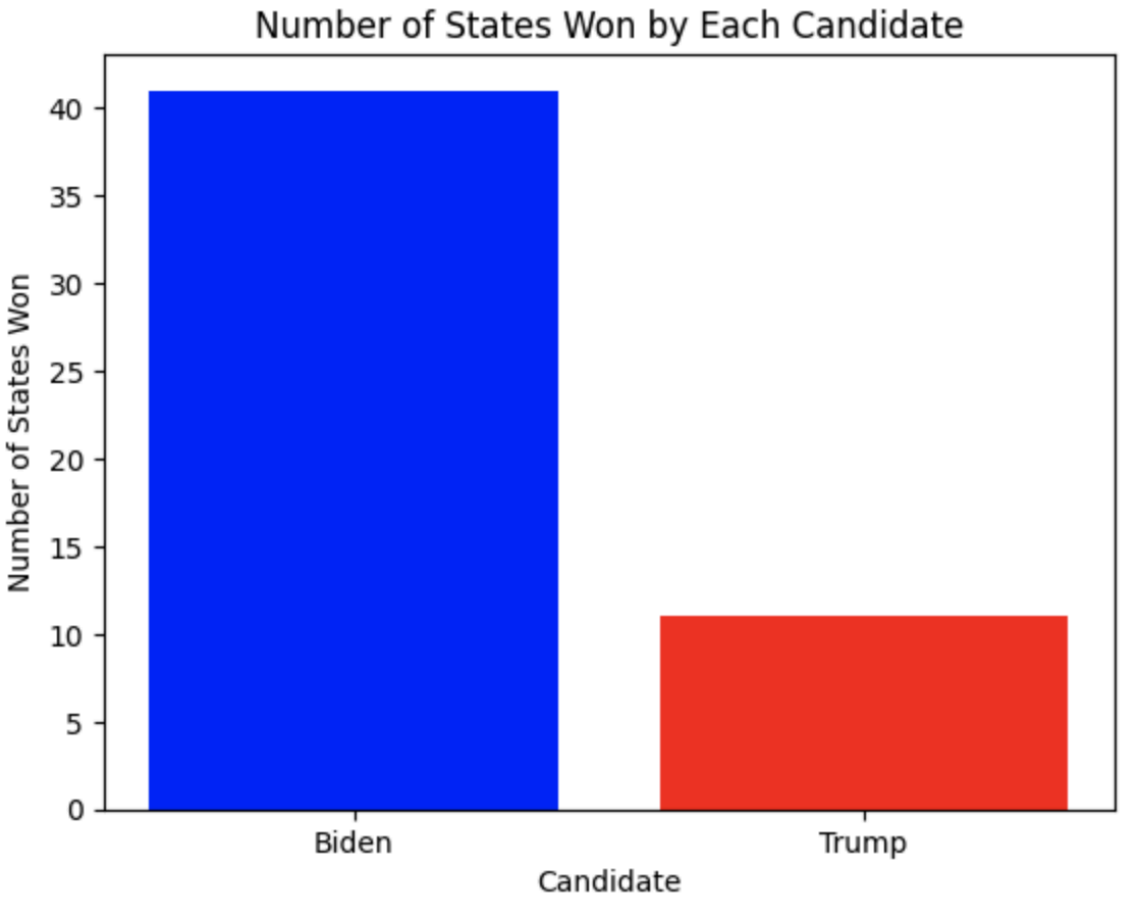


There is quite the spread of discussion regarding Trump, and it may be overlapping the points showing the discussion regarding Biden. But we can see that they are quite equal when it comes to being discussed around the US, though there is a significant amount missing from the Midwestern part of the United States. This could potentially affect the analysis later by incorporating an unintentional bias that will be important to keep in mind.

Sentiment Analysis

In this section, we are ready to conduct our sentiment analysis of the tweets. I iterate through each tweet, analyzing the sentiment polarity of it, and adding this polarity to the end of the dataframe. Following this, I create a subset of the data, grouping the data by state, candidates, and the polarity for each of those candidates. The candidate with the higher polarity value is chosen, and we finish with a dataframe of all the states and their chosen candidates. When we

plot this into a bar graph, we see the clear winner is Joe Biden.

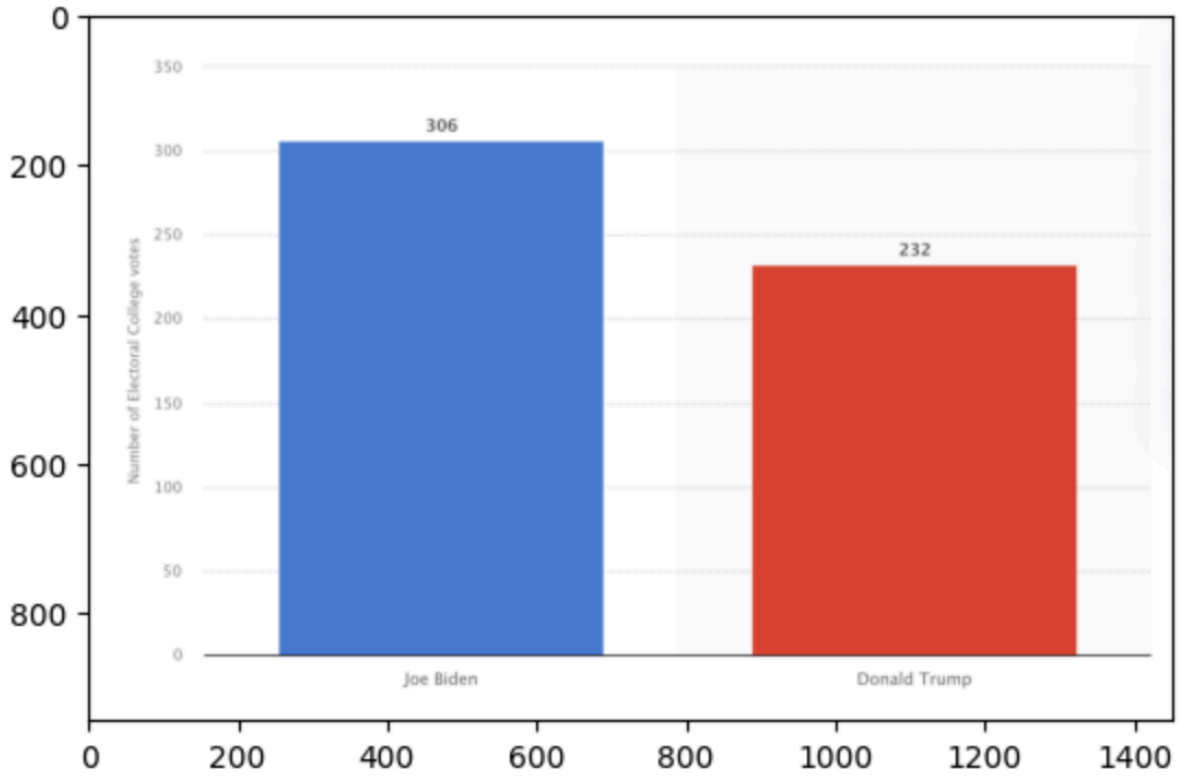


Conclusion

We have seen quite the distribution of discussion around the United States regarding presidential candidates Joe Biden and Donald Trump during the upcoming 2020 election in that time period. We can see from these graphs that the discussion amongst the two candidates is evenly split, though not evenly distributed throughout the country. While there are some gaps in information regarding some of the more Midwestern states, there is still an abundance of information regarding the rest of the country.

After cleaning and processing the data, I utilized TextBlob's sentiment analyzer tool to determine which way a state would vote based on the sentiment of tweets coming from each

state. This involved grouping the dataset with the state and candidate values by the polarity sentiment, followed by keeping the higher of two polarity sentiments, thus the winner of that state. Using the counts of how many would result in Joe Biden or in Donald Trump, I created a bar graph to visualize the results, and we can see that Joe Biden is the clear winner. To verify our results, I obtained a similar graph of the electoral college votes from 2020 and the results were quite similar to my own.



(Published by Statista Research Department, 2024)

Assumptions

While using a social media platform to conduct sentiment analysis to gain an understanding of public sentiment about a candidate can be a good starting point, it is important to remember that Twitter users are not wholly representative of the state. On the contrary, not all

users on Twitter may vote, and not all voters are on Twitter. It is important to take the results of this analysis with this in mind to avoid potential bias. In addition, while sentiment analyzer techniques have improved significantly over time, there are still nuances in text such as sarcasm, irony, or context-dependent expressions that may lead to inaccuracies.

Limitations

Limited data quality, including noise, inaccuracies, and biases inherent in social media data, can also affect the reliability of sentiment analysis results and predictions. This issue can be reduced with proper data preparation and cleaning, but it is hard to avoid completely with user-provided information. Sentiment analysis algorithms may struggle to understand the context of social media posts, leading to misinterpretations or inaccurate sentiment classifications.

There is also the issue that not all users that Tweeted made their locations available, and much of the data had to be removed due to this. It is seemingly random which part of the population made their location unknown. Also, to incorporate a more well rounded analysis of the country's stance on political candidates, the use of other social media platforms may prove to make up for this difference.

Challenges

Addressing biases in social media data, such as demographic biases or algorithmic biases, is critical to ensure fairness and avoid perpetuating existing inequalities. Additionally, ethical considerations regarding user privacy and consent must be carefully addressed. This has been done so by excluding user-specific information, like username and screen name. Achieving accurate predictions of election outcomes based on social media sentiment analysis requires

robust modeling techniques, careful feature selection, and validation to ensure predictive reliability and generalizability.

There was an especially arduous challenge in creating the sentiment analyzer tool. The original goal was to use Transformer's sentiment analyzer tool roBERTa, which is specifically designed in determining the sentiment of tweets. However, this tool proved to be incredibly labor intensive, running for several hours at a point before erroring out for various different reasons. At one point, it even ran for 17 hours nonstop, only to run into an indexing error. I knew the roBERTa tool was an ambitious one, and decided to leave that for a time where there was no time constraint present, and switched to TextBlob.

Future Uses/Additional Applications

A leading reason for the creation of this project was due to the upcoming 2024 election, and how we may be able to predict the outcome based on social media sentiment. At this point, in March 2024, there has not yet been much discussion regarding the election, so this project stands as a starting point for the upcoming year. There is a lot of uncertainty surrounding which party may win this year, and part of the hope with this project is to gain an understanding of how we might begin to predict the winner of the 2024 election using social media sentiment analysis.

An additional application for this project is to be used by either campaigning party to gain an understanding of where their presidential candidate stands amongst the states. They could further use this information to not only understand the sentiment amongst the voters, but why they feel the way that they do, and use this knowledge to reach out to gain their support. This project can be further developed to give presidential candidates an understanding of why they lack support from certain states, and what they may do to mitigate that loss.

Recommendations

There are a few recommendations going forward. First, I recommend revisiting this project as the 2024 election comes closer, and use the knowledge gained from here to conduct further analysis to gain an understanding for where the states stand for the presidential candidates. There are also quite a few ways to improve the accuracy of this project. With the recent takeover of Twitter, now known as X, by a new CEO, there has been a significant drop in its use of discussion, including political discussions. Today, other social media platforms may prove to have a stronger discussion happening, including but not limited to Facebook or TikTok. While there still may be some discussion on X, it utilizing other social media platforms will give a more well rounded analysis of the public's sentiment.

In addition, if the resources and time allow, I recommend the use of the Transformer sentiment analyzer tool roBERTa. It has been trained in over 52 million tweets to understand how to determine its sentiment and has been proved to give a more accurate sentiment analyzer over TextBlob. This will give an overall more accurate representation of the public's sentiment towards one candidate or the other, and allow any team wanting to use this information to act accordingly.

Implementation Plan

In implementing this into campaigning research work to give the campaigning team a better understanding of how the state's opinion stands for the presidential candidate. This can be accomplished by scraping current data from the Twitter API or other social media platform's API and putting it through the same program to be able to gain this level of understanding.

Ethical Assessment

Attempting to predict election results comes with many ethical considerations. To be in compliance with privacy laws protecting user's data, I will ensure to anonymize and randomize user data to protect their privacy. I will be sure to exclude any data that was not given consensually i.e. the user may have opted out of sharing their data, or privatized their profile. Also, because I am obtaining data from social media, there may be a bias present to which demographic of people share their opinions online, and it will be important to disclose this in the final presentation.

Social media platforms are highly susceptible to the spread of misinformation, and I will exercise caution when distinguishing between genuine sentiment and false or misleading information. Lastly, it is important to use the findings of this study responsibly and ethically, and the results will not be used to manipulate public opinion, influence political outcomes, or engage in unethical practices.

10 Questions

Here are 10 questions that I may be asked by the audience:

1. What are the primary objectives of the study?

To gain an understanding into the 2020 election and attempt to predict the winner of the election based on the public sentiment via Twitter discussions.

2. How was the data collected and processed from the Biden and Trump datasets?

The dataset was obtained from a Kaggle dataset. The creator of the dataset stated that they obtained the Tweets from the Twitter API using the ***statuses_lookup*** and ***snsscrape*** for keywords. The tweets range from the beginning of 2020 up to 11/8/2020.

3. What methods were employed to extract state information from user locations in the datasets?

If the user allowed it, the user's location was part of the data extracted from the Twitter API. From here, I only included the tweets where the user's location was present, excluding any unknown locations.

4. How were missing or incomplete data handled during the analysis?

As this study was location based, I dropped any missing user location. I also excluded any tweets that did not originate from within the United States, as this study is solely focused on those in the United States.

5. What specific metrics or criteria were used to determine the candidate affiliation of each tweet?

The candidate affiliation was predetermined by the Kaggle dataset creator. Their method was to distribute them into either the Biden or Trump dataset by whoever's name was stated in the tweet, no matter the polarity or sentiment of the tweet.

6. How were the total numbers of tweets per state calculated for both Biden and Trump?

By creating a stacked bar graph that showed the distributions of tweets for both Biden and Trump. The specific calculations of the tweets were not done, but more so a visualization to confirm that the distribution of the tweets for the two candidates was equal.

7. Were any statistical tests conducted to compare the distribution of tweets per state between Biden and Trump?

No statistical tests were conducted to compare the distribution, but a visualization of the distribution of tweets per state was done to show this distribution. It showed that there was a largely disproportionate amount of tweets coming from the large three states, California, New York, and Texas, most likely due to the larger population.

8. What are the implications of the findings for understanding public sentiment and engagement on social media during the election period?

During the election period, tensions were higher and it was hard to gain an understanding of where the public stood in regards to one presidential candidate or the other. The implications of understanding public sentiment and engagement on social media would allow us to gain an

understanding of where each state stands in regards to the candidates. It provides insight, and with further analysis, could provide the candidates with ideas on how to improve their campaigning techniques to have a stronger reach on the voters.

9. How might the results of the study contribute to our understanding of the political landscape and dynamics across different states?

The results will allow us to understand how the voters in each state feel in regards to a presidential candidate. This is especially important for swing states, where they may vote for either party. These states are incredibly important to presidential candidates, as gaining the favor of these states will swing the vote in their direction.

10. What potential limitations or biases should be considered when interpreting the results, and how might they impact the validity and generalizability of the findings?

It is important to keep in mind that there was a disproportionate amount of tweets coming from the three major states: California, New York, and Texas. There were big gaps in the Midwestern part of the United States, as demonstrated in one of the graphs. This implies that the results of the sentiment analysis for those particular states is not wholly representative of the population. It is also important to remember that those who are active on Twitter are not always the ones who vote, and the ones who vote are not always active on Twitter.

Appendix

```
biden.describe()
```

	likes	retweet_count	user_followers_count	lat	long
count	181137.000000	181137.000000	1.811370e+05	181137.000000	181137.000000
mean	12.753706	2.964684	1.141703e+04	37.548902	-93.370002
std	530.538091	80.771621	1.030344e+05	4.936310	16.765426
min	0.000000	0.000000	0.000000e+00	13.450126	-161.755833
25%	0.000000	0.000000	1.220000e+02	34.053691	-100.445882
50%	0.000000	0.000000	6.200000e+02	39.516223	-90.199838
75%	2.000000	0.000000	2.699000e+03	40.712728	-77.727883
max	165702.000000	20615.000000	5.750841e+06	65.538963	145.639196

```
trump.describe()
```

	likes	retweet_count	user_followers_count	lat	long
count	213263.000000	213263.000000	2.132630e+05	213263.000000	213263.000000
mean	9.333907	2.367481	1.036617e+04	37.618705	-93.616049
std	274.404234	68.294247	9.551262e+04	4.893897	16.834602
min	0.000000	0.000000	0.000000e+00	13.450126	-162.597762
25%	0.000000	0.000000	1.250000e+02	34.098003	-100.445882
50%	0.000000	0.000000	6.210000e+02	39.551150	-92.561787
75%	1.000000	0.000000	2.589000e+03	40.712728	-77.727883
max	74084.000000	20491.000000	5.747472e+06	71.387113	144.757551

References

Manch Hui. 2020, November). US Election 2020 Tweets, Version 3. Retrieved March 31, 2024 from https://www.kaggle.com/datasets/manchunhui/us-election-2020-tweets.

Published by Statista Research Department. (2024, February 22). *2020 presidential election: Results U.S. 2020*. Statista. https://www.statista.com/statistics/1184537/2020-presidential-election-results-us/