Capstone project 3: Sentiment Analysis on the US Election Candidates

1.0 BACKGROUND

The 2020 US election happened on the 3rd November 2020 and the resulting impact to the world will no doubt be large, irrespective of which candidate is elected!

Twitter was a top social app that collected huge size of comments on both candidates positively and negatively. Ironically and debatably, Twitter suspended candidate Trump's account and some conservative voices and groups, and therefore a lot of Trump's followers left Twitter. In this case, the Twitter's comments might not truely reflecting the voices regarding both candidates. But it is a decent dataset for NLP practice for sentiment analysis.

2.0 GOAL

The goal of this project is to carry out EDA and sentiment analysis based on the datasets provided. Hopefully to check if there is any correlation between the election result and the text messages here.

3.0 DATA

- Link to data: https://www.kaggle.com/manchunhui/us-election-2020-tweets (https://www.kaggle.com/manchunhui/us-election-2020-tweets)

Tweets collected, using the Twitter API statuses_lookup and snsscrape for keywords, with the original intention to try to update this dataset daily so that the timeframe will eventually cover 15.10.2020 and 04.11.2020. Added 06.11.2020 With the events of the election still ongoing as of the date that this comment was added, I've decided to keep updating the dataset with tweets until at least the end of the 6th Nov. Added 08.11.2020, just one more version pending to include tweets until at the end of the 8th Nov.

Columns are as follows:

- created_at: Date and time of tweet creation
- tweet_id: Unique ID of the tweet
- tweet: Full tweet text
- likes: Number of likes
- retweet_count: Number of retweets
- source: Utility used to post tweet
- user_id: User ID of tweet creator
- user_name: Username of tweet creator
- user_screen_name: Screen name of tweet creator
- user_description: Description of self by tweet creator
- user_join_date: Join date of tweet creator
- user_followers_count: Followers count on tweet creator
- user_location: Location given on tweet creator's profile

- lat: Latitude parsed from user_location
- long: Longitude parsed from user_location
- city: City parsed from user_location
- country: Country parsed from user_location
- state: State parsed from user_location
- state_code: State code parsed from user_location
- collected_at: Date and time tweet data was mined from twitter*

Kaggle kernels referenced

- https://www.kaggle.com/manchunhui/us-presidential-election-sentiment-analysis (https://www.kaggle.com/manchunhui/us-presidential-election-sentiment-analysis)
- https://www.kaggle.com/tkubacka/a-story-told-through-a-heatmap (https://www.kaggle.com/tkubacka/a-story-told-through-a-heatmap)
- https://www.kaggle.com/harikrishna9/who-won-in-us-elections-2020-according-to-tweets (https://www.kaggle.com/harikrishna9/who-won-in-us-elections-2020-according-to-tweets)

4.0 Method

- Step 1: Exploratory Data Analysis:

Besides the text data (tweets) in both Trump and Biden's dataset, the location of the twitter users are also collected. The dataset is a good one to carry out EDA to understand the dataset, like the location of the users, the devices used, and also the languages used.

- Step 2: N-gram and wordcloud:

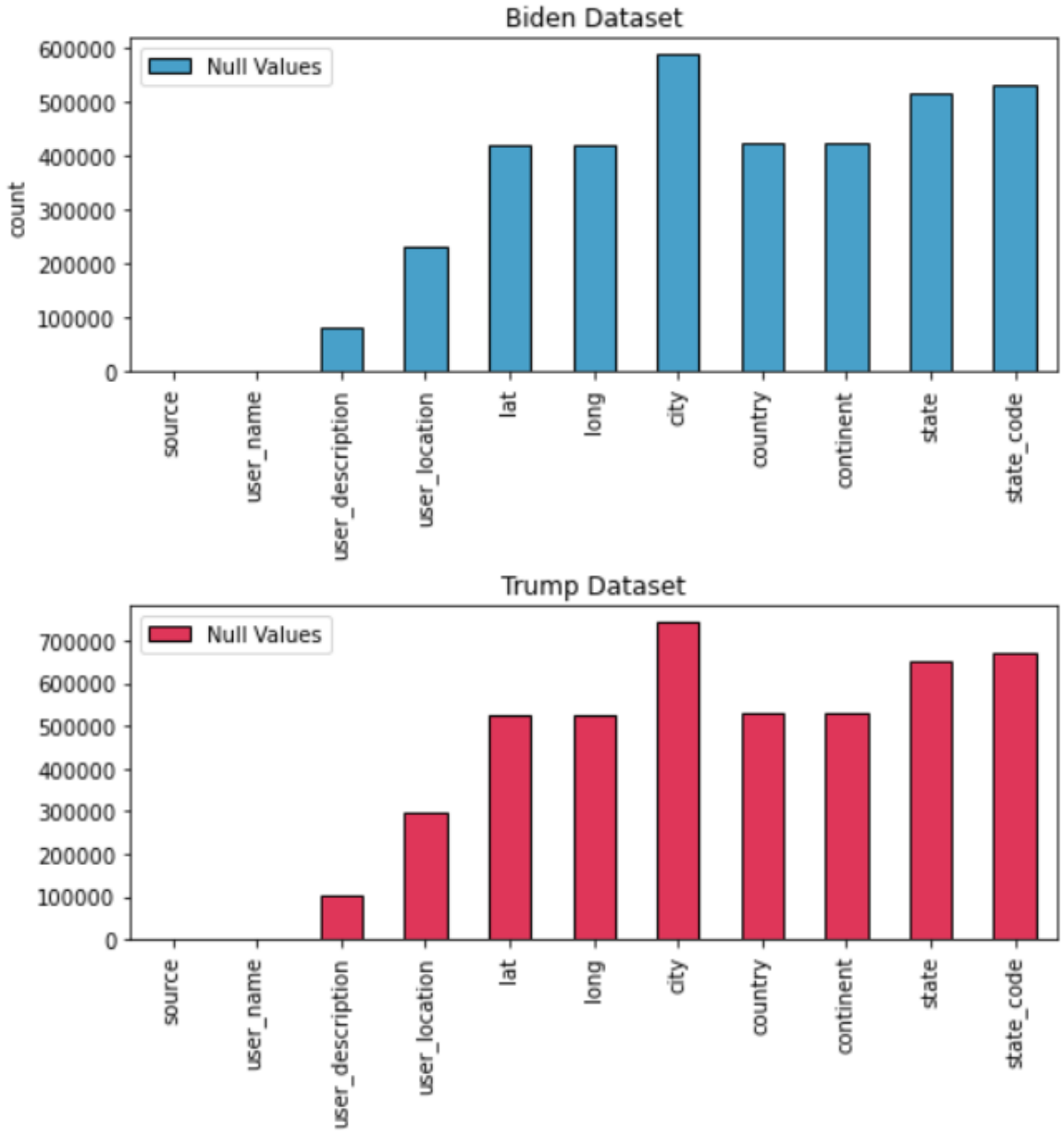
The N-gram method and the wordcloud will give a direct view of high frequency word(s) in both Trump and Biden's tweets.

- Step 3: Sentiment Analysis (VADER):

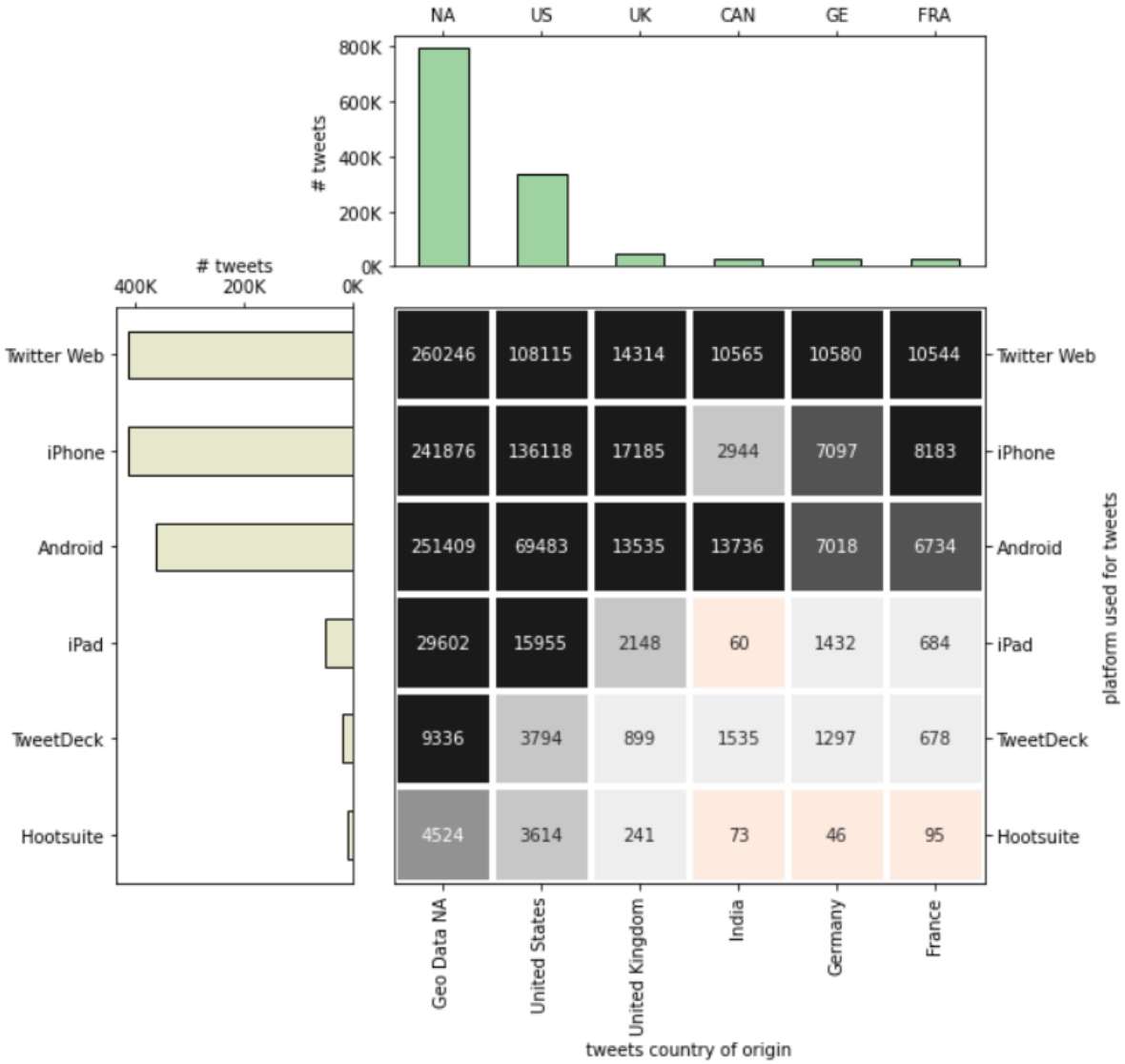
To perform the Sentiment Analysis I will be using VADER (Valence Aware Dictionary and sEntiment Reasoner) package, which is a lexicon and rule-based sentiment analysis tool that is specifically tuned to sentiments expressed in social media!

5.0 Data Cleaning and EDA

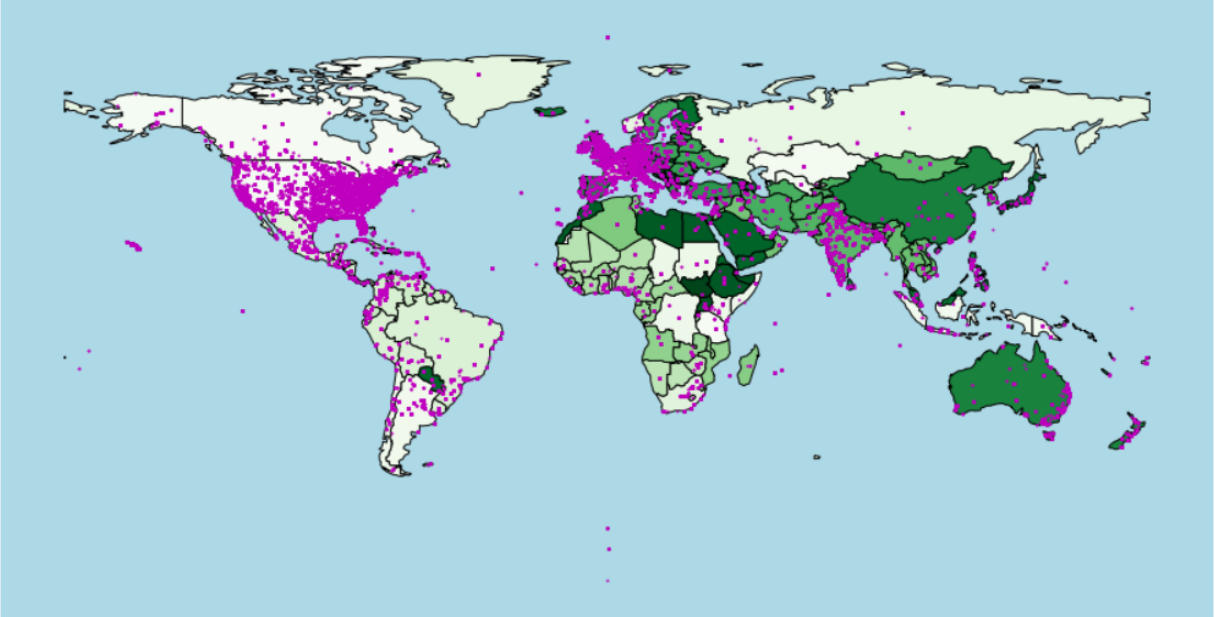
The bar plot below shows a summary plot of numerical features in both Trump and Biden dataset. The missing values of different features exist for both trump and biden dataframe. In general, trump dataset has more missing values for each feature.



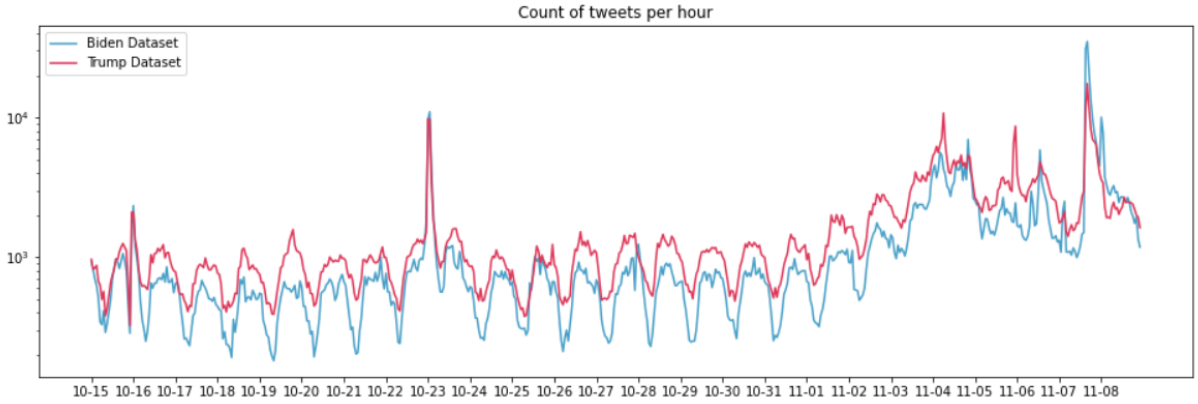
This matrix plot is plotting the devices used for tweeting comments versus the countries. It is clear that we have a lot of missing values. Beyond that, US is the top country having the most comments. Twitter web, iphone, and android are leading devices for making tweet comments.



Since the dataset has geo-spatial data. Geopandas library was used to plot the user locations. This Geoplot here gives a visible view of all the tweets comments coming from. It is clear that the whole world is very interested in this US election as the future president will have big impact on global politics.



This plot gives an alternative view of trump and biden comments amount from the publishing data perspective. It is clear that generally trump is more popular than biden from October to November.



6.0 N-grams and Sentiment Analysis

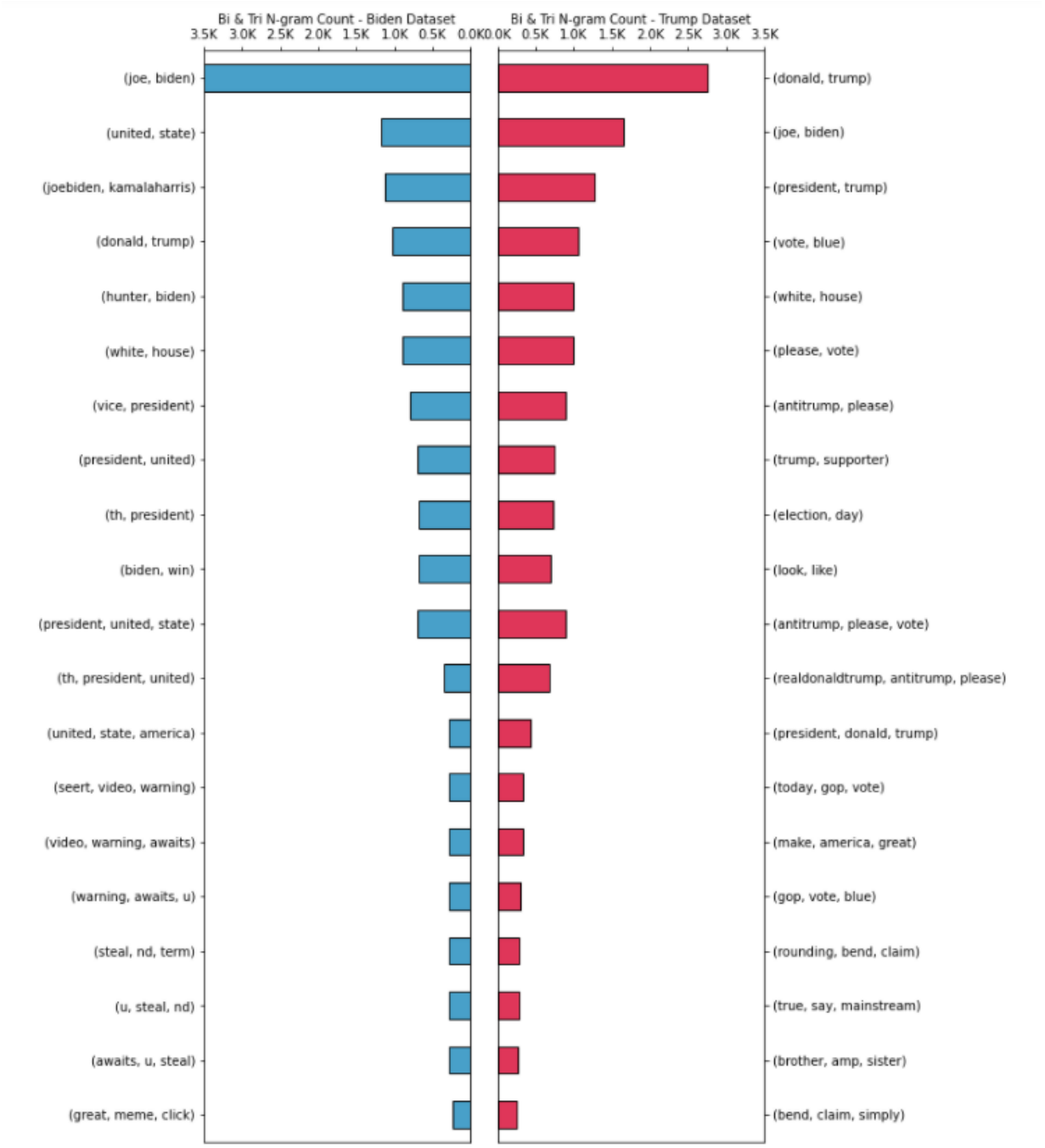
For the sentiment analysis purpose, the data from United States of America is only used.

6.1 N-grams

Before the N-gram analysis we first must clean the tweets to remove stopwords, strings with "http" etc and then lemmatize the words.

N-grams are contiguous sequences of n-items in a sentence. N can be 1, 2 or any other positive integers, although usually we do not consider very large N because those n-grams rarely appears in many different places.

The bar chart above shows the most common Bi and Tri N-gram's in both trump and biden text data published in USA only. The N-gram's show a clear relation to the upcoming election and each respective dataset seems to be related to each of the presidential candidates.



I also created wordcloud to show the most frequennt words in the text data.

6.2 Sentiment Analysis (VADER)

To perform the Sentiment Analysis I will be using VADER (Valence Aware Dictionary and sEntiment Reasoner) package, which is a lexicon and rule-based sentiment analysis tool that is specifically tuned to sentiments expressed in social media!

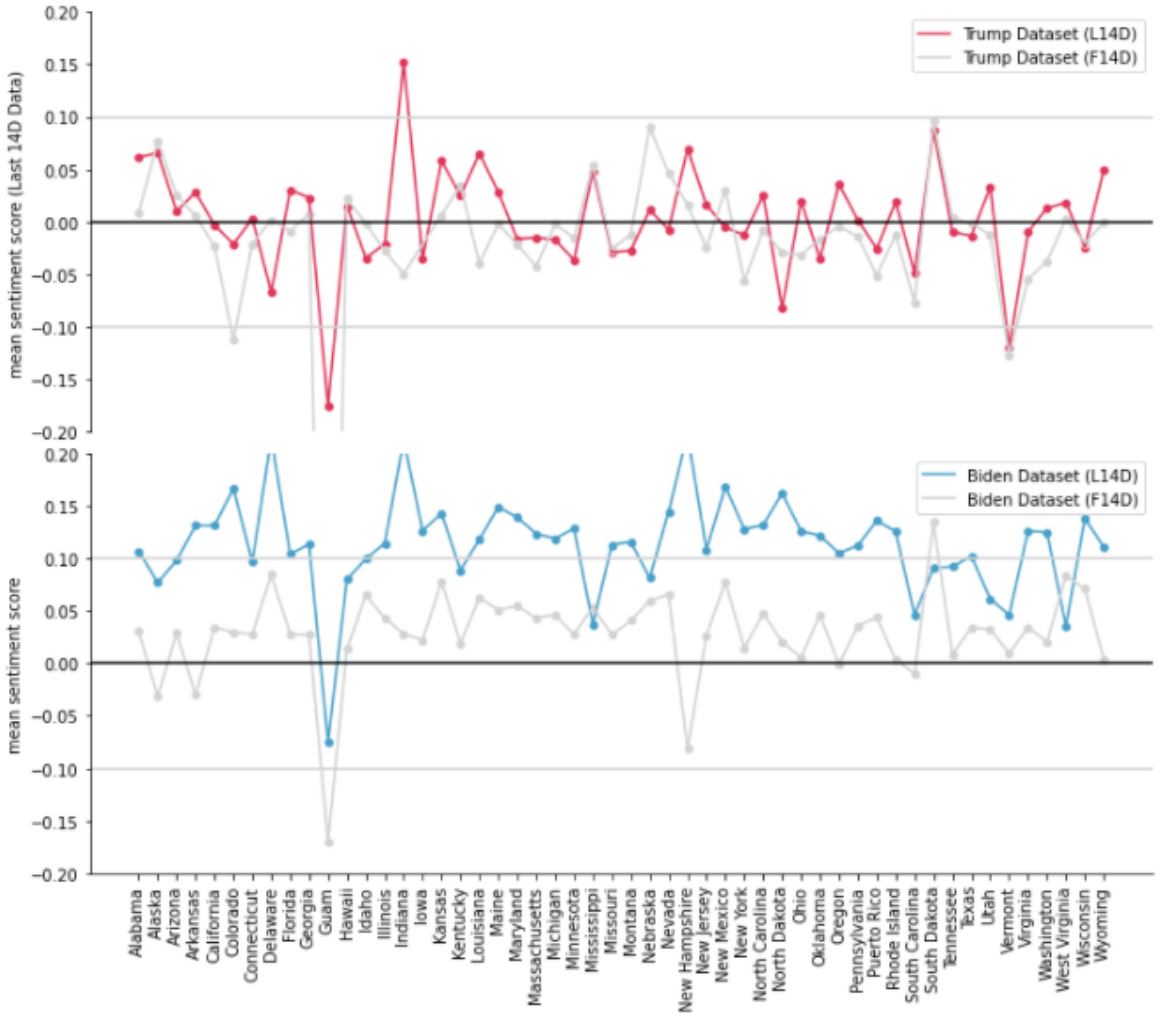
SentimentIntensityAnalyzer function is used to predict the sentiment score for each tweet comment. The feature name is "compound". The range of "compound" is from -1 to 1. For the score between -0.1 and 0.1 is set to be neutral comments. If score is above 0.1, the sentiment is positive; if lower than -0.1, the sentiment is negative.

7.0 Sentiment Analysis Results

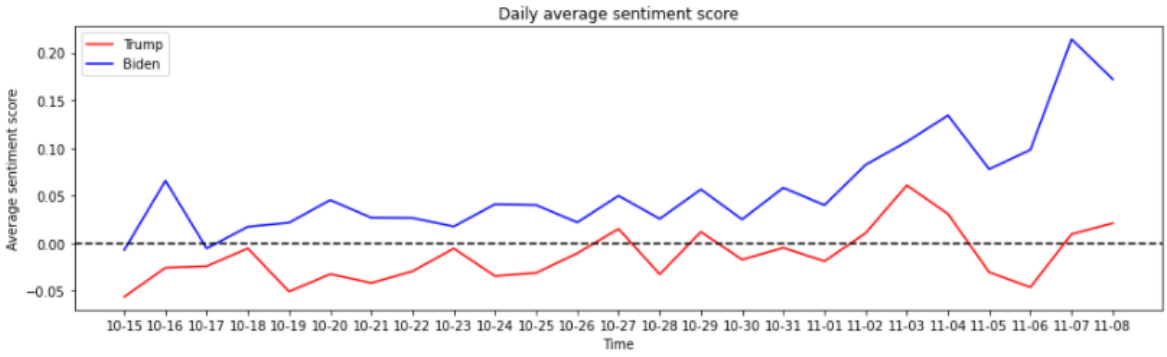
The below plot shows the average sentiment score for all electing states for both Trump and Biden. The most recent 14 days is Labelled L14D on the charts and the first 14 days Labelled F14D and light grey on the charts for each state.

It should be noted that any sentiment score between -0.1 and 0.1 is considered "Neutral".

The results seems to show a large number of states are trending to a “Positive” sentiment score for the democratic candidate from the previous more “Neutral” sentiment. Whereas most states are still largely "Neutral" for the republican candidate.



Another interesting angle to look at the sentiment analysis result is to plot the daily average score for both candidate along with time. From the daily average sentiment score plots of Trump and Biden, it is clear that Biden has higher sentiment score over time. For majority time, Biden has positive score whiel trump has negative score.

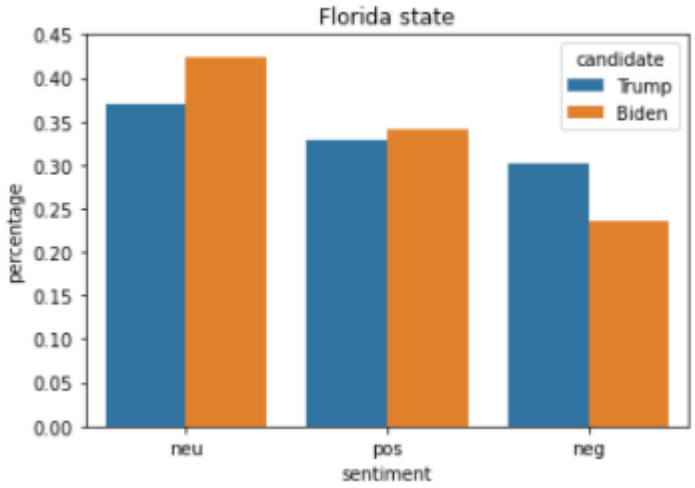


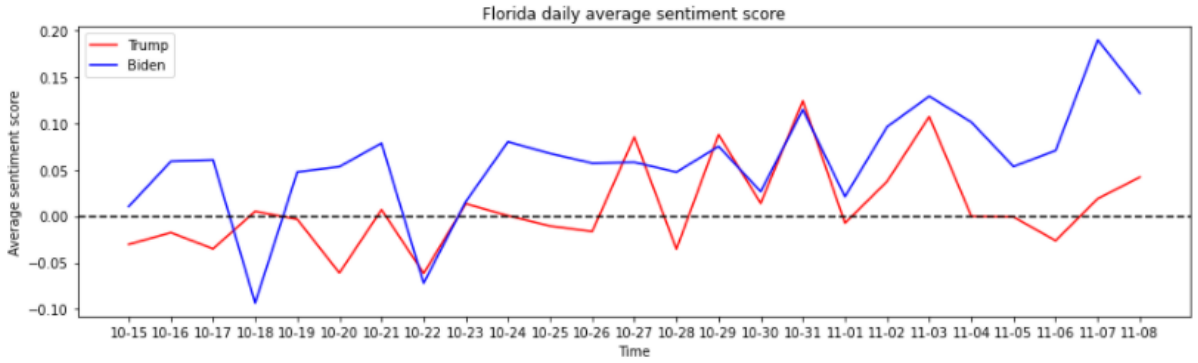
The sentiment analysis was performed only on data that had geo-data originating from the "United States of America" to try to ascertain the sentiment in each respective dataset and therefore each presidential candidate. When reviewing sentiment at the state level as we approached the election date a large number of states were trending to a “Positive” sentiment score for the democratic candidate from the previously more “Neutral” sentiment. Whereas most states are still largely “Neutral” for the republican candidate. This trend is correlatable when viewing the sentiment analysis from a date perspective.

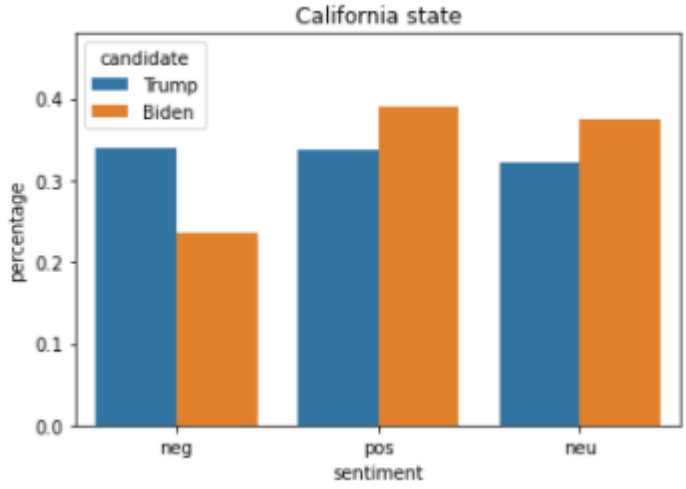
7.1 Florida and California

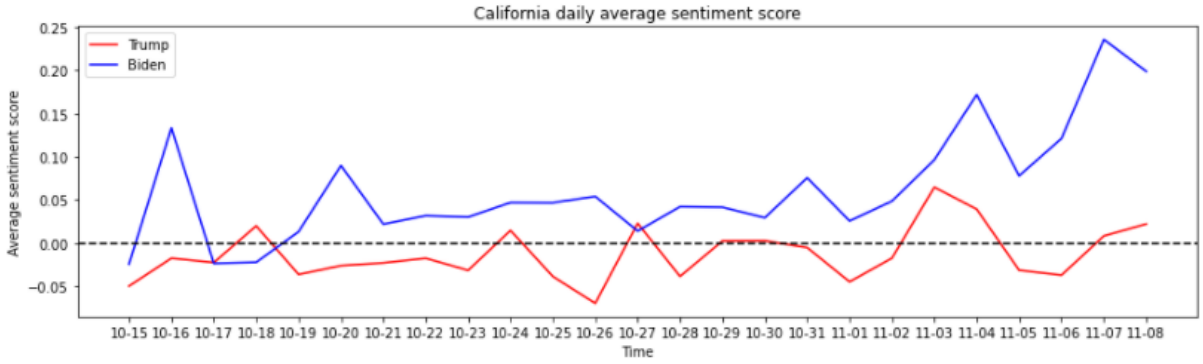
Florida (republican state) voted for Trump and California (democrats) voted for Biden. I am going to use both of them to verify whether the data sample and the conclusion from the data sample is trustable or not.

The sentiment analysis suggests that Biden should win Florida. However this is not the reality. Therefore, the dataset is not representative to the whole US electing population.









8.0 Conclusion and Future Recommendations

Conclusion

The twitter text data give a sample to analyze which candidate has better chance to win the 2020 US presidential election. Trump has a larger dataset than Biden. It is safe to say that Trump is more popular than Biden. Based on the work above, it is clear that Biden generally has higher sentiment socre in most states. Biden won the 2020 election.

People might say that twitter data reflects the election reality. However, when look at Florida and California state, the Florida result is opposite with the twitter data analysis. From the statistical standpoint, it is safe to say that twitter data is not a good sample to represent the whole popupation.

Recommendations

1. Geopandas library is not easy to install. A good suggestion is to build a virtual environment to install Geopandas and its dependencies.
2. From the statistical standpoint, the dataset is not a good sample to understand the whole US voting behavies and results. But it is still a good dataset for sentiment analysis.

9.0 Credits

Thanks to Rajib Biswas for being an amazing Springboard mentor.

In [ ]: