Capstone Project 3: Sentiment Analysis on US Election 2020

Tingting Zhang

09/10/2021

Content:

- Problem Statement
- Data Source and Reference
- Method
- Exploratory Data Analysis (EDA)
- N-grams and Wordcloud
- Sentiment Analysis (VADER)
- Summary
- Recommendation

Problem Statement:

- Twitter is a top social app that collected huge size of comments on both Republican and Democrats candidates during the 2020 US presidential election.
- A good study would be carrying out sentiment analysis on twitter comments about both candidates and check if twitter comments is a good sample for the whole US election population.

Data Source and Reference:

- Link to data:

https://www.kaggle.com/manchunhui/us-election-2020-tweets

- Features:

1. created_at: Date and time of tweet creatio;
2. tweet_id: Unique ID of the tweet;
3. tweet: Full tweet text;
4. likes: Number of likes;
5. retweet_count: Number of retweets;
6. source: Utility used to post tweet;
7. user_id: User ID of tweet creator;
8. user_name: Username of tweet creator;
9. user_screen_name: Screen name of tweet creator;
10. user_description: Description of self by tweet creator;
12. user_followers_count: Followers count on tweet creator;
13. user_location: Location given on tweet creator's profile;
14. lat: Latitude parsed from user_location;
15. long: Longitude parsed from user_location;
16. user_join_date: Join date of tweet creator;
17. city: City parsed from user_location;
18. country: Country parsed from user_location;
19. state: State parsed from user_location;
20. state_code: State code parsed from user_location;
21. collected_at: Date and time tweet data was mined from twitter

- Kaggle kernels referenced:

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

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.

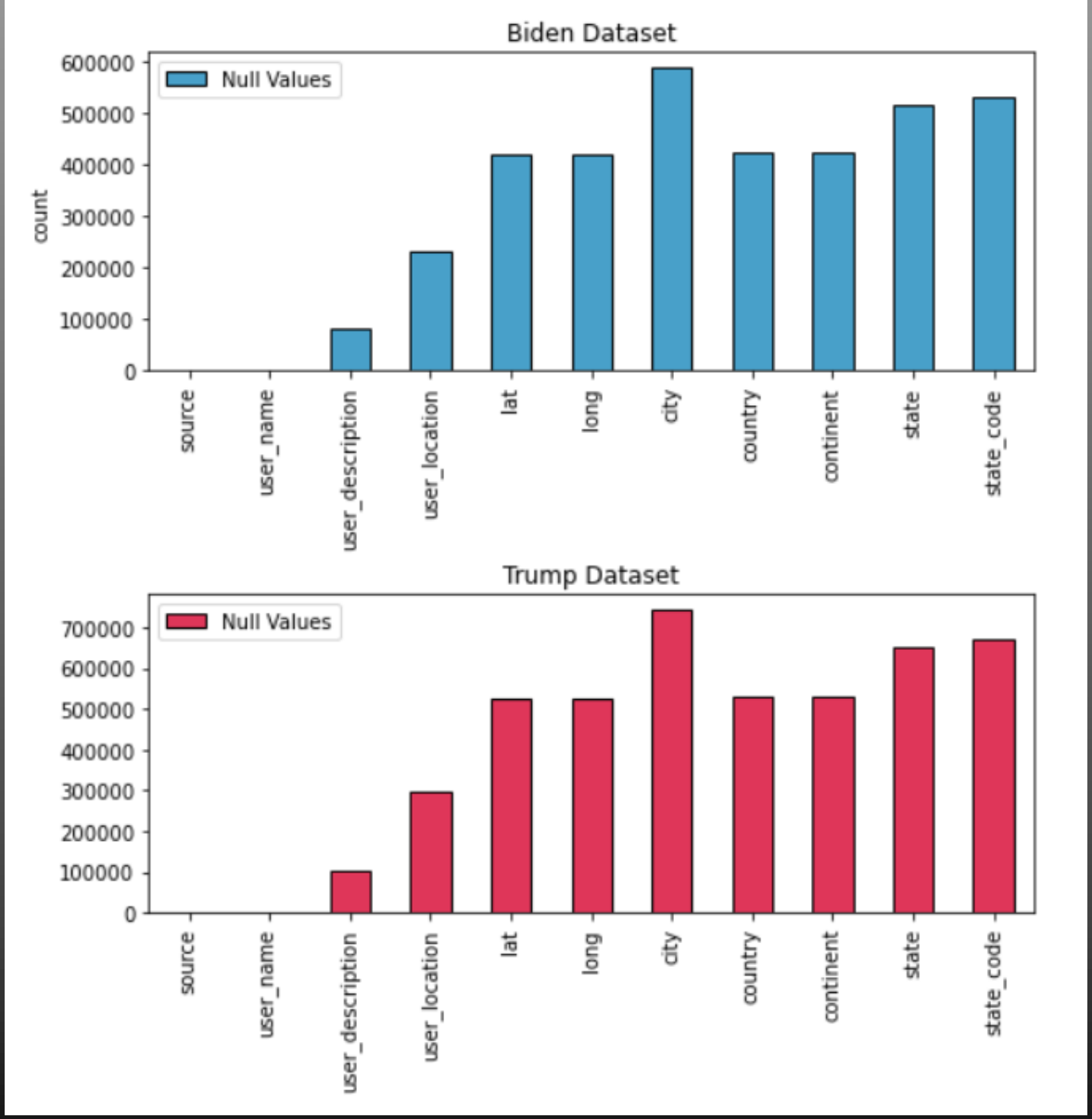
- Step 2: N-gram and wordcloud:

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):

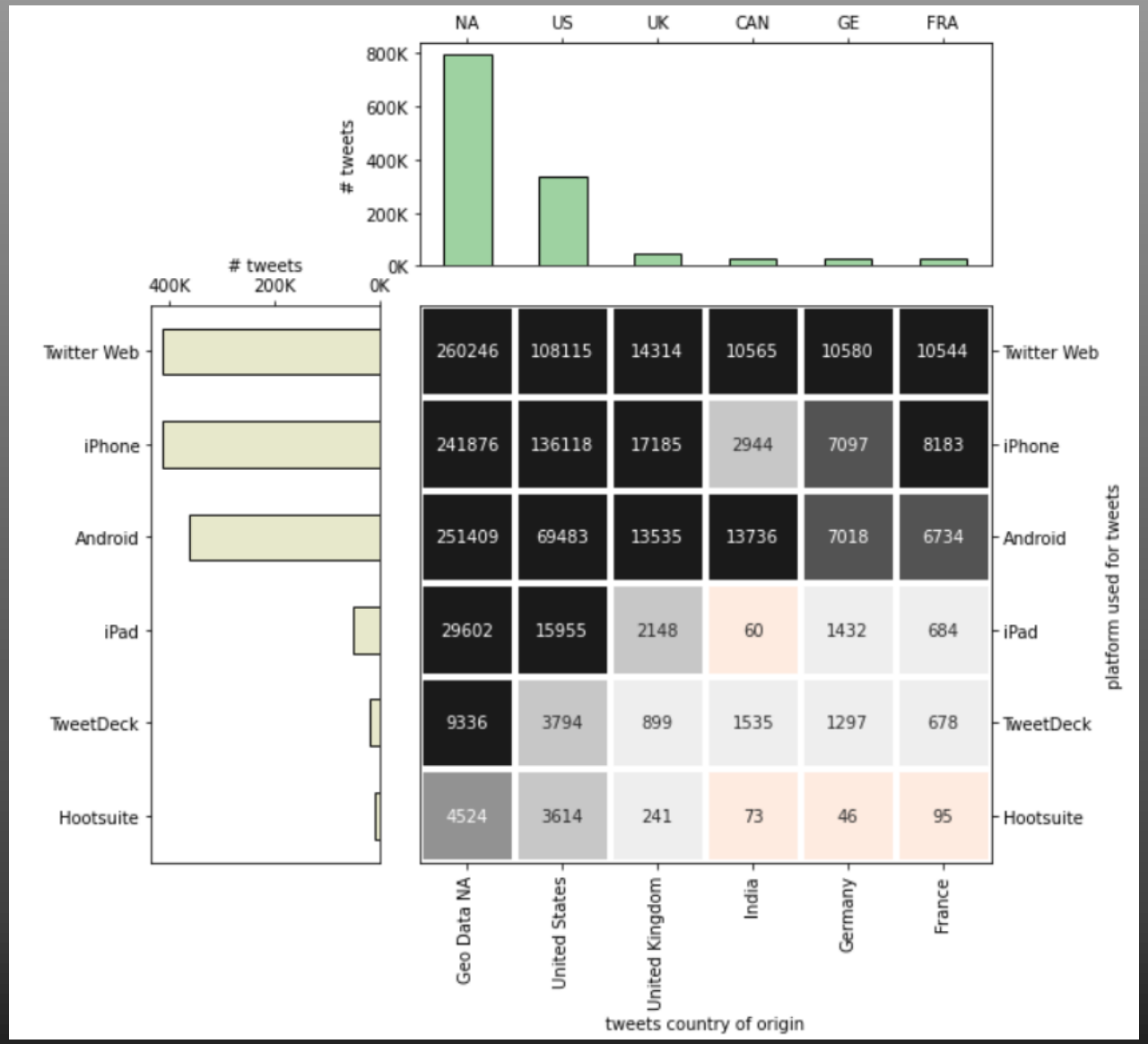
To perform the Sentiment Analysis, I will be using VADER (Valence Aware Dictionary and sEntiment Reasoner) package, which is a lexicon and rule-based sentiment analysis tool that is specifically tuned to sentiments expressed in social media!

Exploratory Data Analysis:



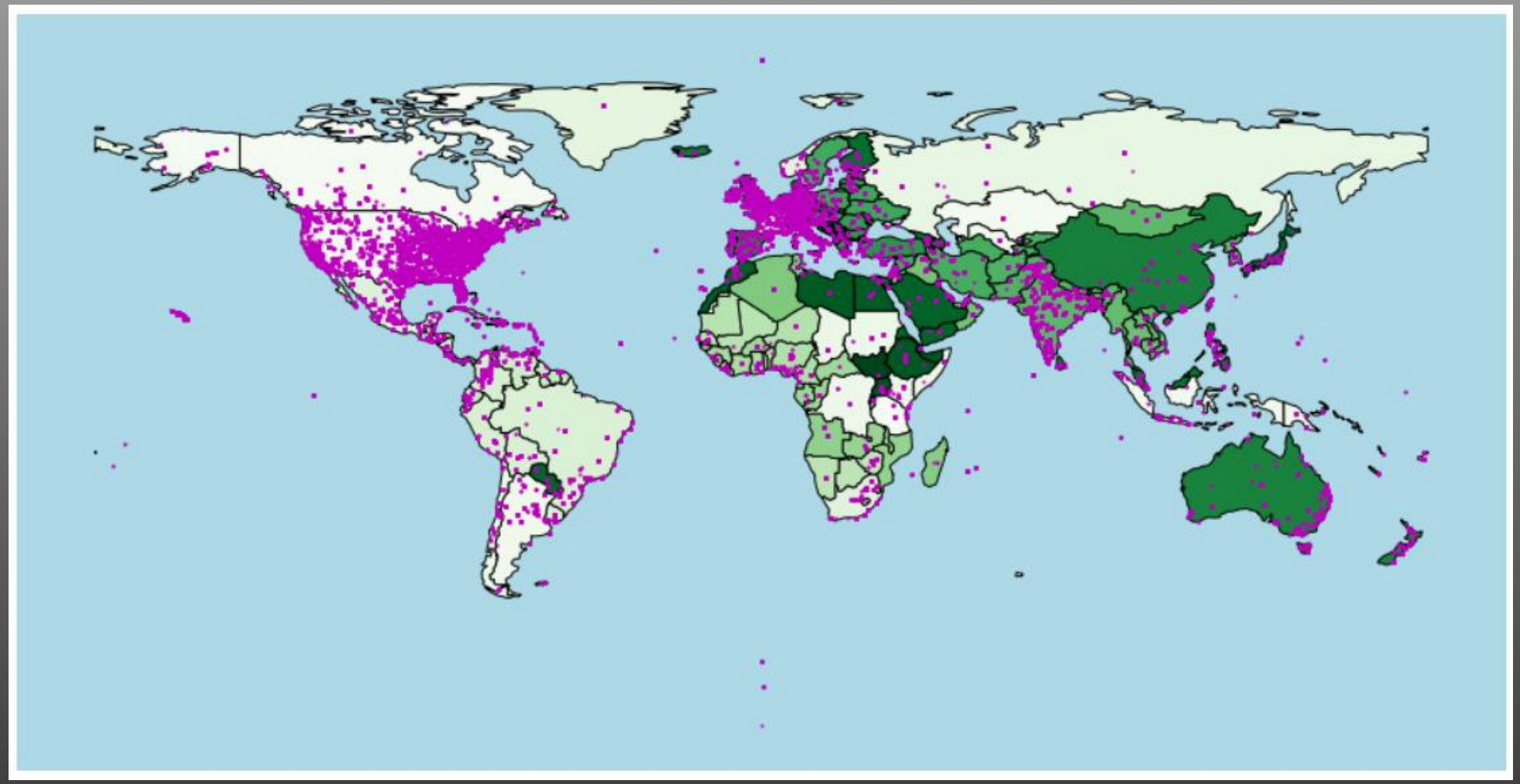
- The bar plot below shows a summary plot of numerical features in both Trump and Biden dataset. The missing values of different features exist for both Trump and Biden dataframes.
- In general, trump dataset has more missing values for each feature.

Exploratory Data Analysis:



- The matrix plot here is a great way of plotting categorical data.
- 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.

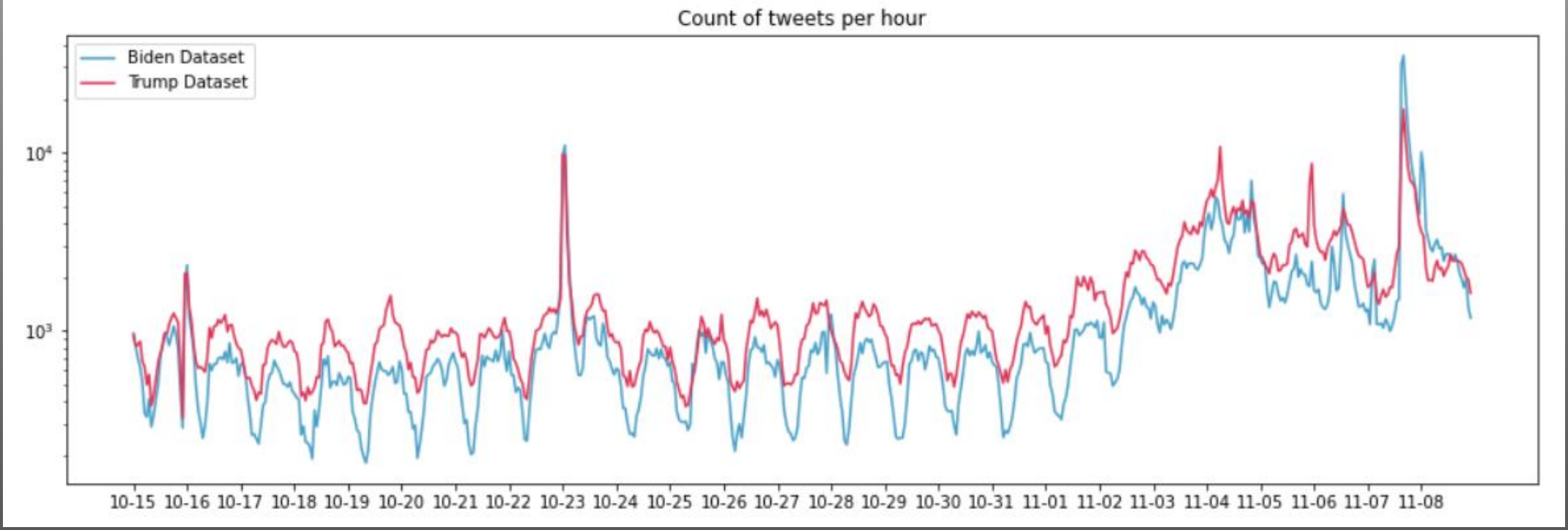
Exploratory Data Analysis:



The user location is plotted with Geopandas library. Majority of data is from United States of America.

Europe countries also show great interest in 2020 US presidential election.

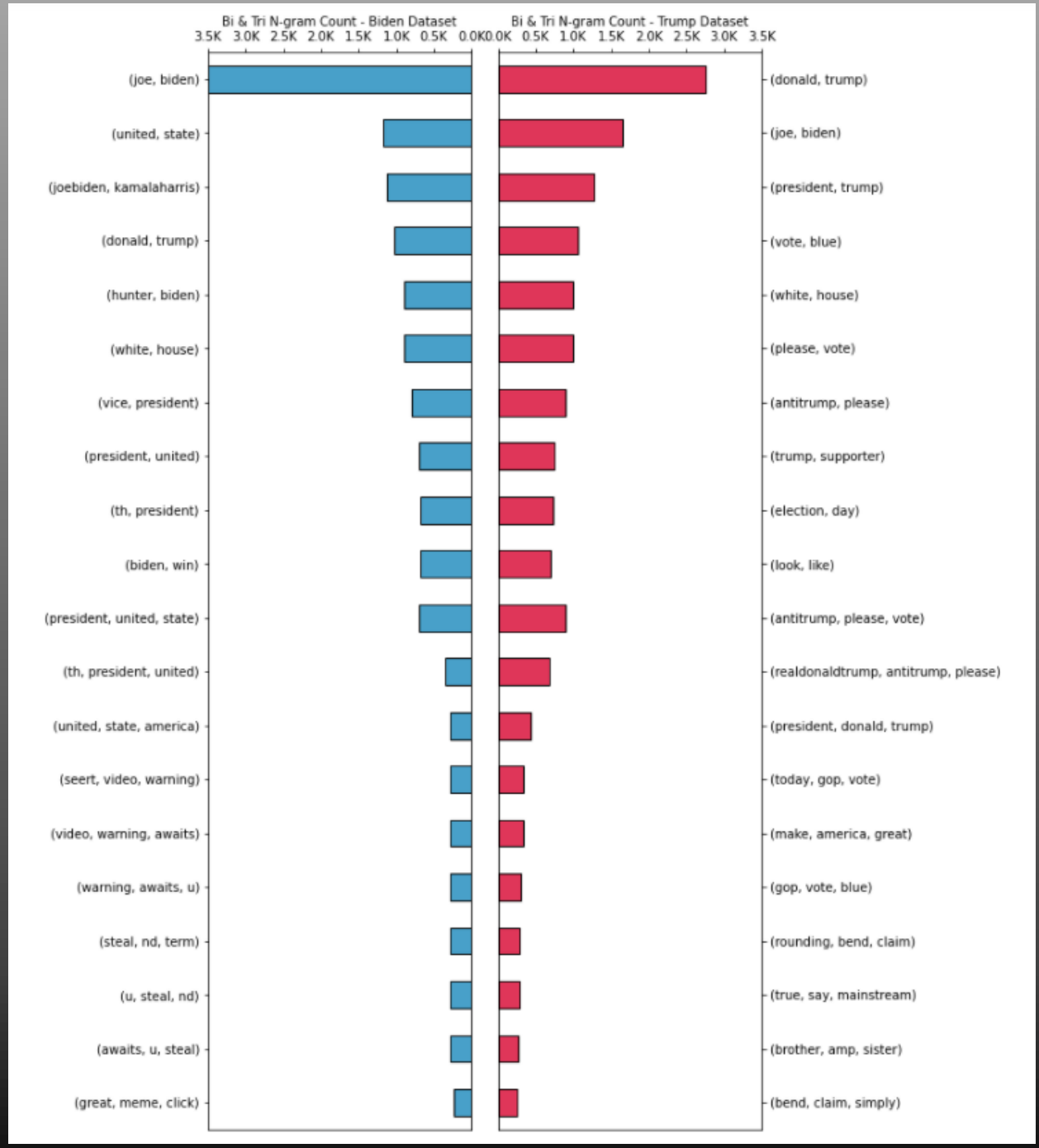
Exploratory Data Analysis:



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

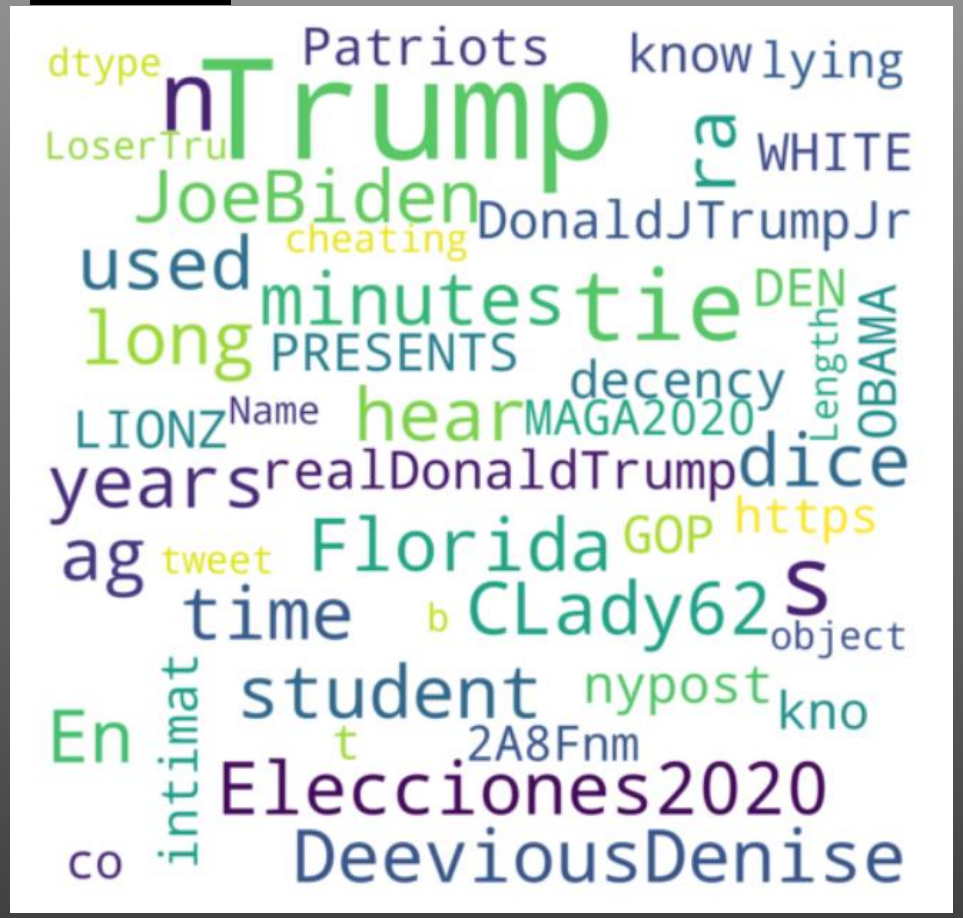
N-grams

- N-grams are contiguous sequences of n-items in a sentence. N can be 1, 2 or any other positive integers.
- The bar chart above shows the most common Bi and Tri N-grams 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.

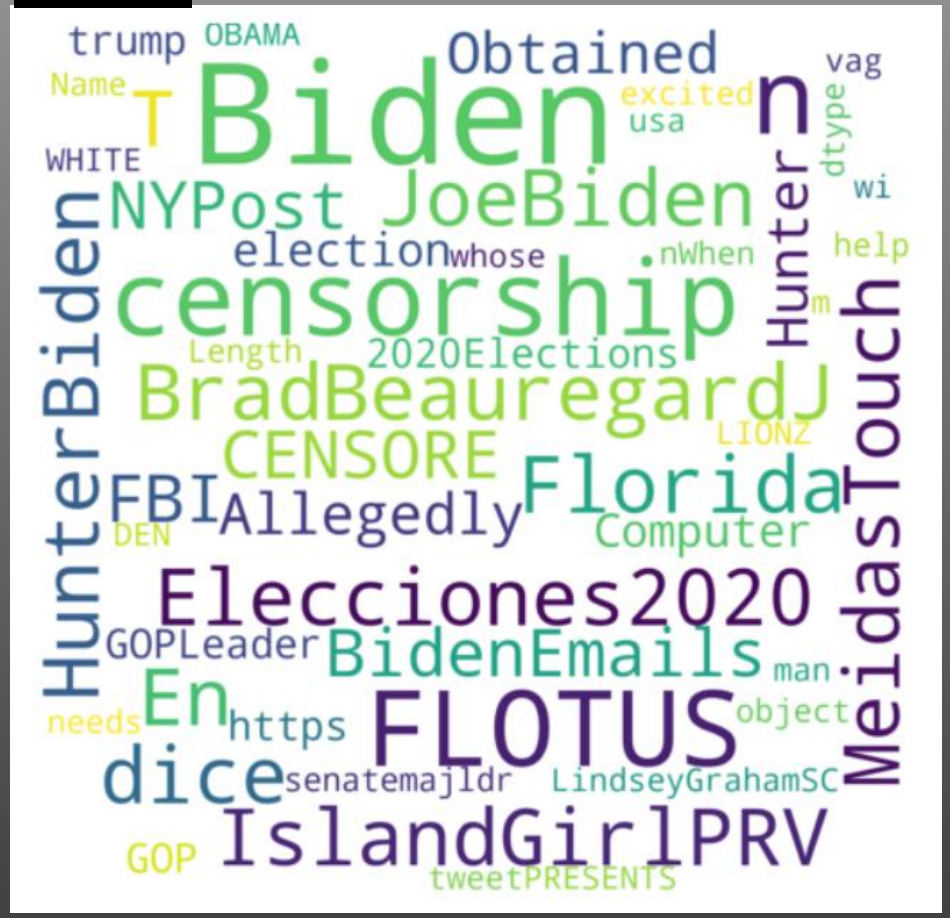


Wordcloud

Trump



Biden



The wordcloud is a good visualization for high frequency words in both candidates' tweets.

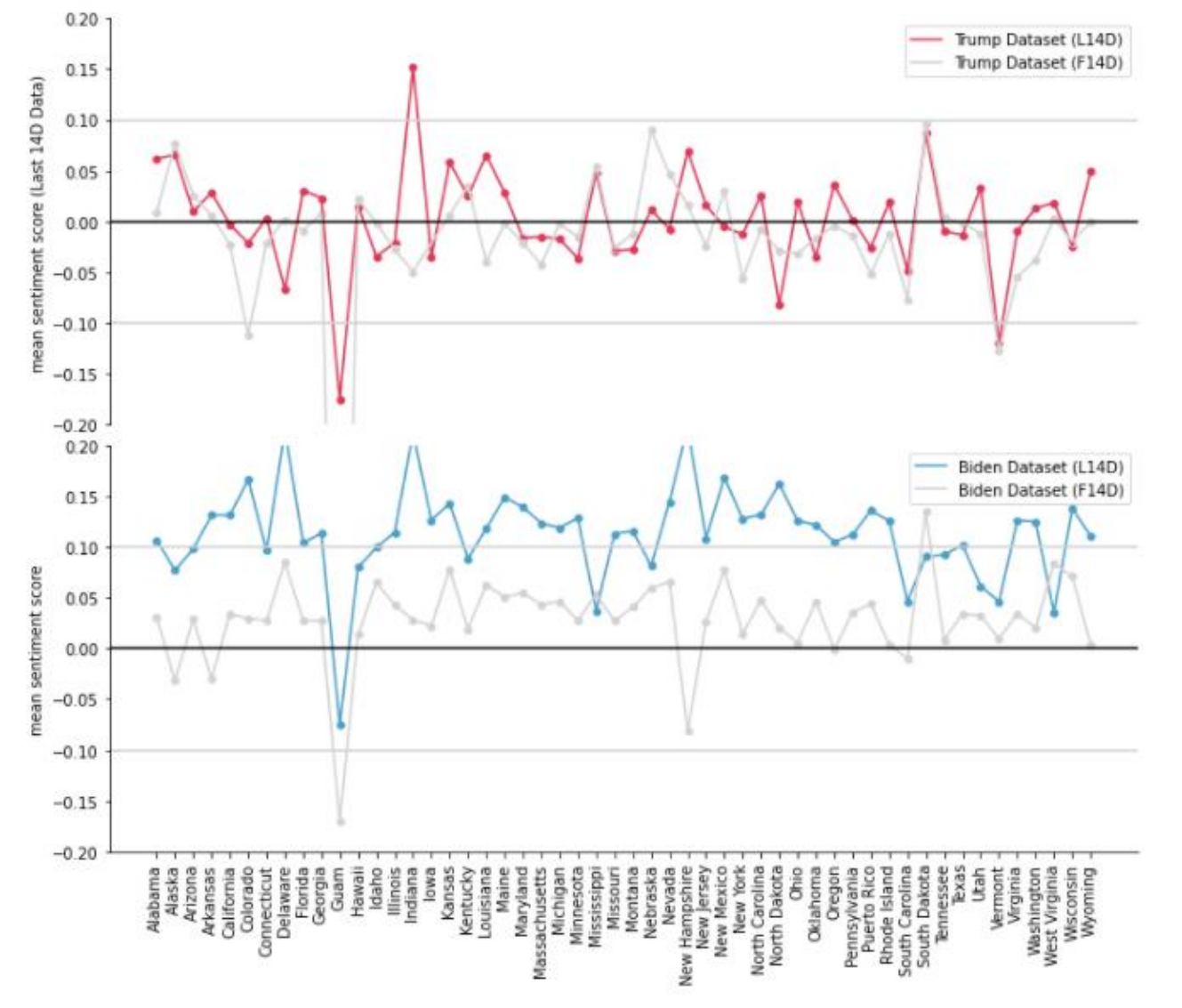
Sentiment Analysis (VADER)

```
# Obtain sentiment scores for both datasets
sid = SentimentIntensityAnalyzer()
biden['VADAR']=sentiment(biden['tweet'])
trump['VADAR']=sentiment(trump['tweet'])
biden['compound']  = biden['VADAR'].apply(lambda score_dict: score_dict['compound'])
trump['compound']  = trump['VADAR'].apply(lambda score_dict: score_dict['compound'])
trump['sentiment']  = trump['compound'].apply(lambda x: 'pos' if x > 0.1 else ('neg' if x < -0.1 else 'neu'))
biden['sentiment']  = biden['compound'].apply(lambda x: 'pos' if x > 0.1 else ('neg' if x < -0.1 else 'neu'))
```

- VADER (Valence Aware Dictionary and sEntiment Reasoner) package is used to perform the Sentiment Analysis. It 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.

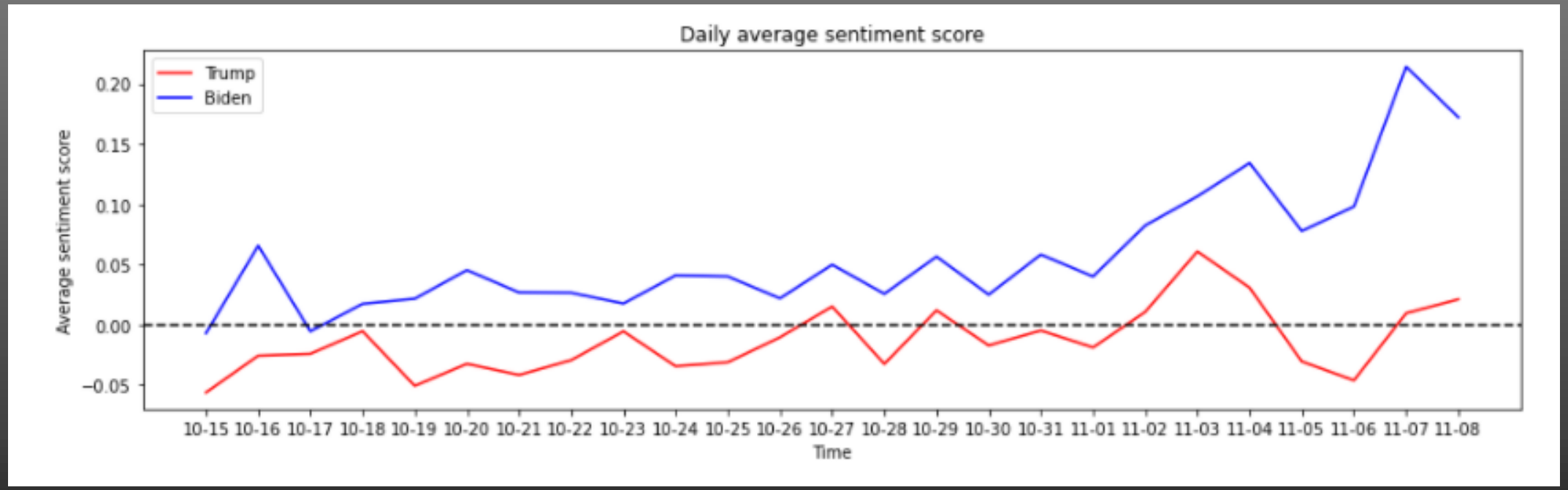
Sentiment Analysis (VADER)



- The 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 that most 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.

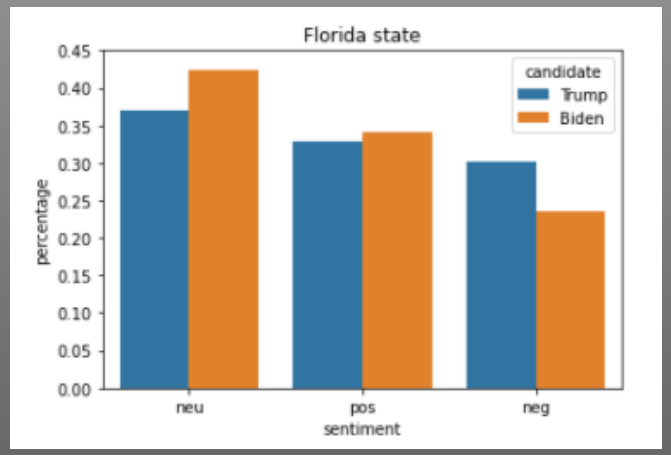
Sentiment Analysis (VADER)

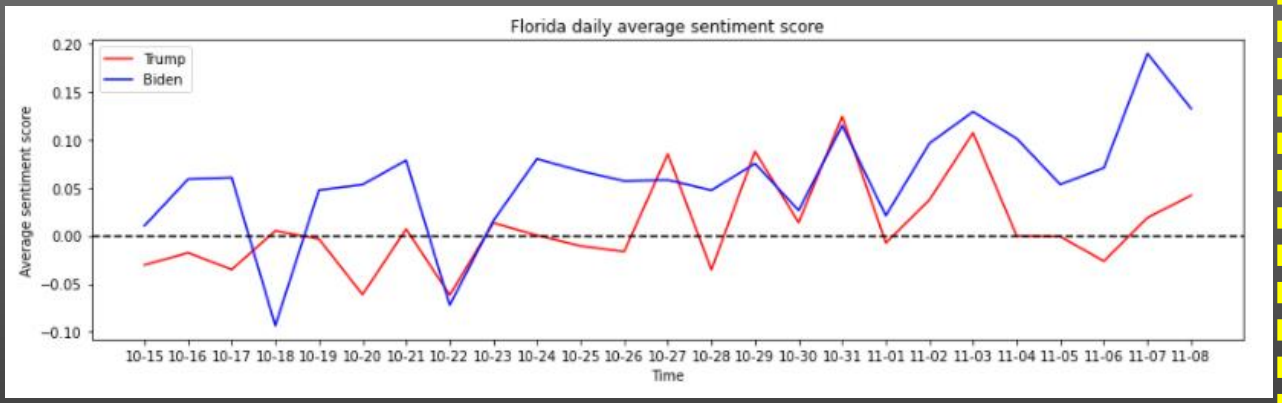
- The below plot is the daily average score for both candidate along with time.
- It is clear that Biden has higher sentiment score over time. For majority time, Biden has positive score while trump has negative score.



Florida and California

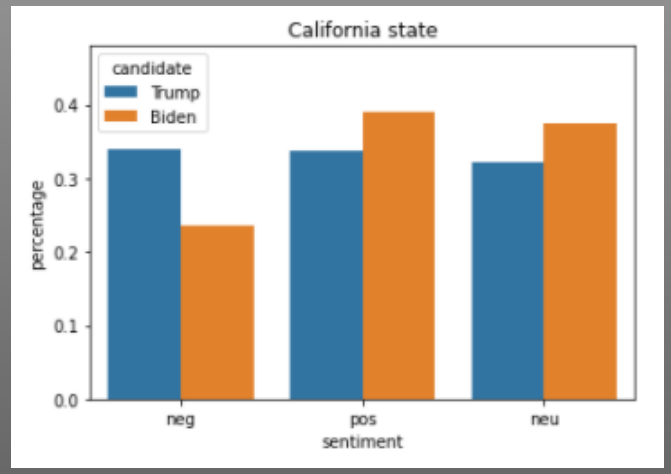
Florida (Trump won)

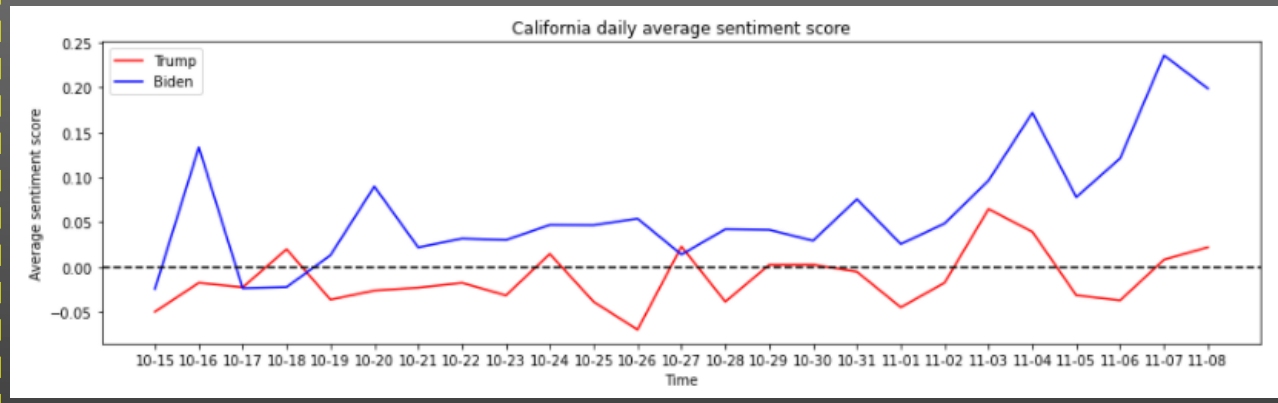




Florida is a republican state during 2020 presidential election. However, the tweets show that Biden received more positive comments than Trump.

California (Biden won)





California is a deep blue state. And the tweets show that Biden is way more positive than Trump in this state.

Summary:

- The twitter social app provides a good dataset for NLP practice of sentiment analysis to understand how twitter users react to both 2020 US presidential candidates.
- Generally, Trump is more popular than Biden as his tweets more than Biden's.
- The sentiment analysis shows that Biden has more positive comments than Trump, which matches the 2020 election outcome.
- However, a close look at key states indicates that twitter users in US is Biden prone no matter they are in republican state or democrat state.
- Therefore, from the statistical standpoint, it is safe to say that twitter data is not a good sample to represent the whole US election population.

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 behaves and results. But it is still a good dataset for sentiment analysis.

Thank you!!!

Thanks to Rajib Biswas for being an amazing Springboard mentor!