CS 529 - Introduction to Machine Learning

Program 3 – Sentiment Analysis

Submitted By: **Submitted To:**

Vasman Kaur Prof. Lydia Tapia

UNM ID: 101892375

Section 1:

The Naïve Bayes is based on calculating the probability of the words and then making predictions for the sentiment. Given a sentence in training dataset we take the words and make the positive and negative word pool based on the class label given (class label being positive or negative sentiment) and calculate the probability for negative and positive class which is number of positive sentiment words divided by the total positive and negative sentiment words and the probability of negative class also in the same way. This is called prior probability p(c) where c is positive or negative class. Now for each sentence, here each tweet the probability of each word occurring in a class is calculated for both positive and negative class. This is called the likelihood (p(wi|c)). These probabilities (prior probability and likelihood for each word in a sentence) are multiplied and calculated (positive and negative separately for each sentence and for each word in the sentence) and for a sentence if its positive probability is greater than the negative probability it is classified as positive and vice versa.

The formula for naïve bayes used was:

Class pred = max (P(c)*p(wi|c))

Training Accuracy:

1.) Accuracy when the data was raw i.e., data was not cleaned and had URLs, usernames, hashtags, emojis etc.

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
93.32035053554041
```

Fig 1. Accuracy for raw data

2.) Accuracy when just the URL was removed.

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
92.0253164556962
```

Fig 2. Accuracy for data with no URL

3.) Accuracy when URL and username was removed.

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
83.54430379746836
```

Fig 3. Accuracy for data with no URL and username

4.) Accuracy for data when URL, Username, Hashtags were removed.

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
82.35637779941577
```

Fig 4. Accuracy for data with no URL, username, and hashtag

5.) Accuracy when the data was completely cleaned, i.e., URLs, names, hashtags, and punctuation were removed.

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
97.4001947419669
```

Fig 5. Accuracy for data with no URL, username, hashtag, and punctuation

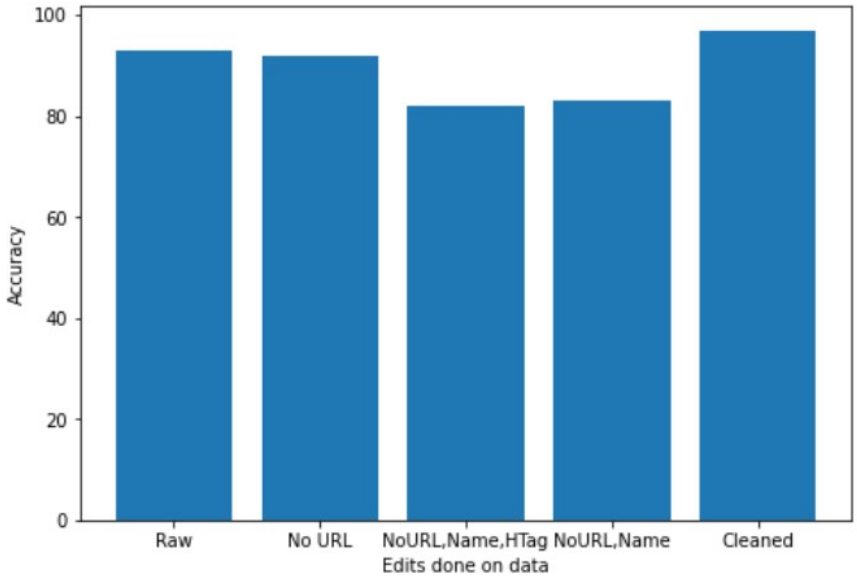

Fig 6. Accuracy vs edits on the data for Training dataset

From the above plots it is seen that the accuracy was high when the data was raw and nothing was edited out the accuracy was high and as the data was removed the accuracy decreased, removing URLs, names and hashtags showed the accuracy decrease from 93% to 82%. Till here the outcome showed that the names and hastags and urls had some importance in sentiment analysis as removing them was decreasing the accuracy and the assumption was made that if the emojis or punctuations which are used in the emojis are removed the accuracy will further decrease as emojis are seen to have more sentimental value or shows emotions clearly. But when from the data after removing urls, names, hastags the punctuations were also removed the accuracy when to 97% so the assumption made about emojis was wrong.

Testing Accuracy:

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
51.26637554585153
```

Fig 7. Accuracy for Testing dataset

For testing dataset the accuracy on the data after completely cleaning it(no url, no name, no hashtag, no punctuation) is seen to be 51%. There was not much change in in accuracy value even after selective cleaning(removing just the url, or name and url or juts the name or name and hashtag and url or other combinations).

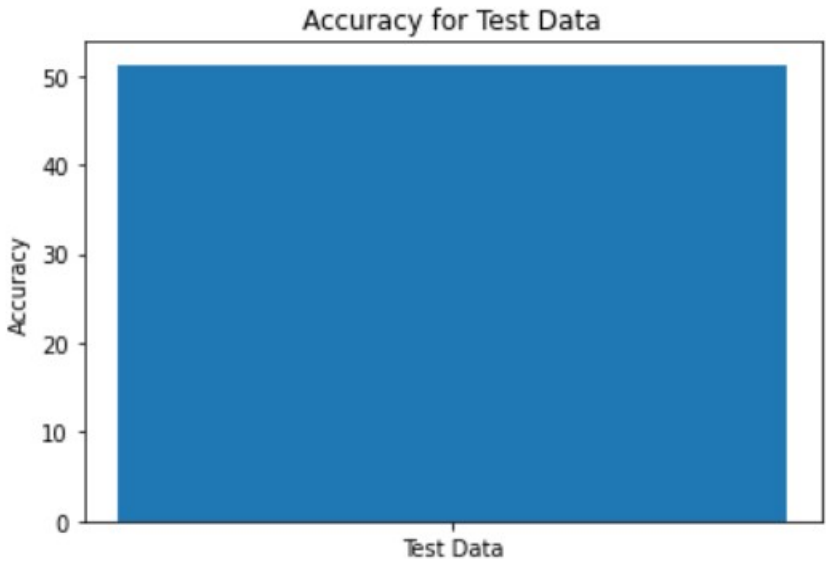

Fig 8. Accuracy for Testing dataset

Extra (Experiment only): I additionally tried one more thing which was using testing dataset as training dataset (which is not supposed to be like this because testing dataset is for testing the model but just out of curiosity) and the accuracy for training (actually testing) dataset then was pretty impressive. All the values were near 99%. So, another conclusion from this is maybe the data needed to clean a little more or maybe the word pool from the training dataset had not much of a variety for the probabilities to be calculated in such a way so as to make correct predictions for the sentences. Below is the plot for that.

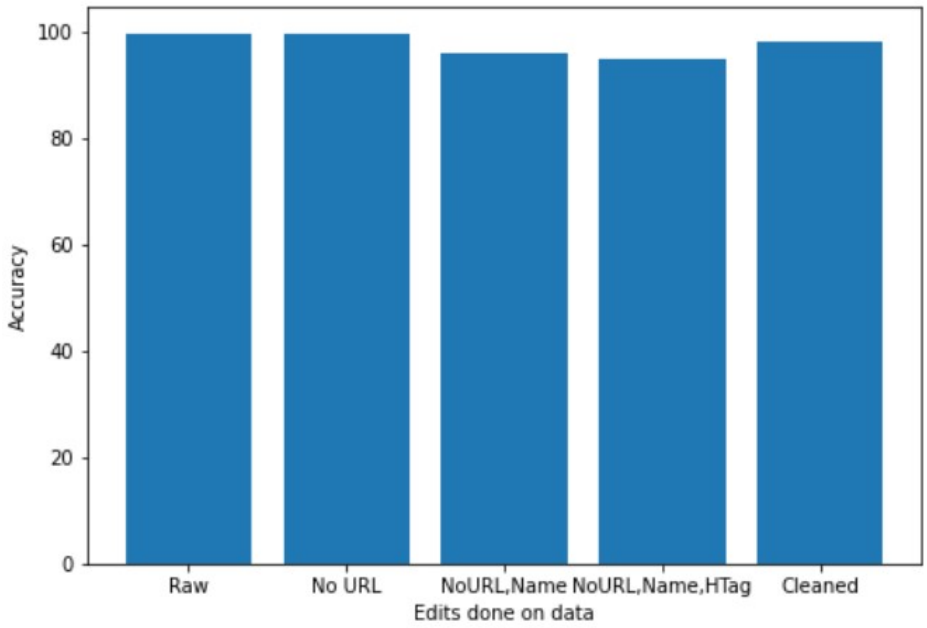

Fig 9. Accuracy Experiment

Section 2: Election Tweets

The analysis for election tweets: using the naïve bayes model described in section1 the election tweets were classified into negative and positive tweets. Below is the outcome graph. From total 1693 predictions 600 were categorized as negative and the remaining 1093 as positive. Which shows overall positive sentiment was higher.

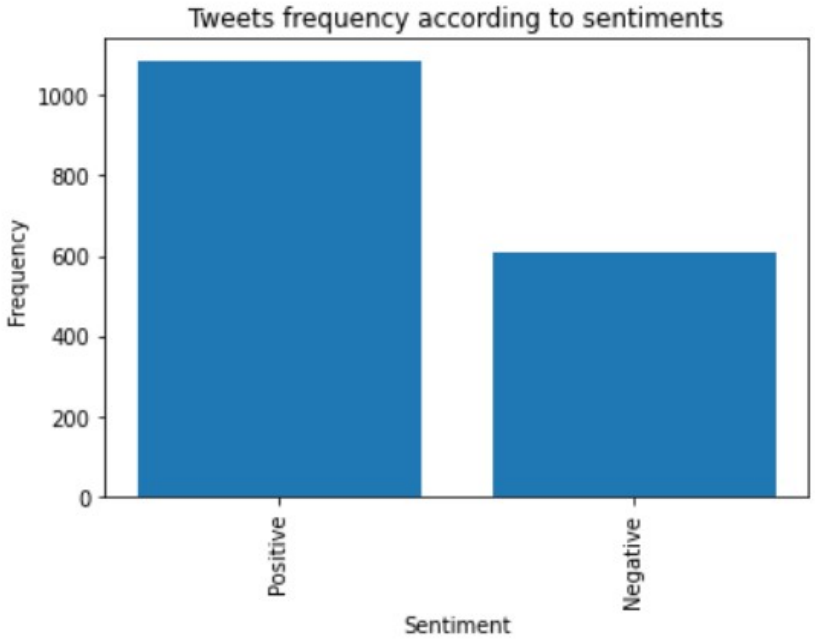

Fig 10. Tweets according to sentiment

From the positive and negative tweets classification it was further analyzed that who among the two candidates were the tweets for. It is seen that overall Donald Trump has higher number of tweets in both the cases, negative and positive.

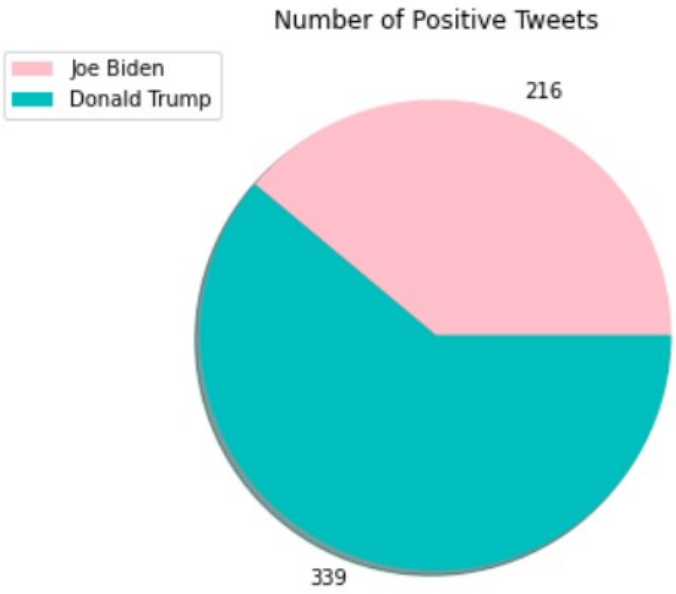

Fig 11. Positive Tweets classified based on candidates.

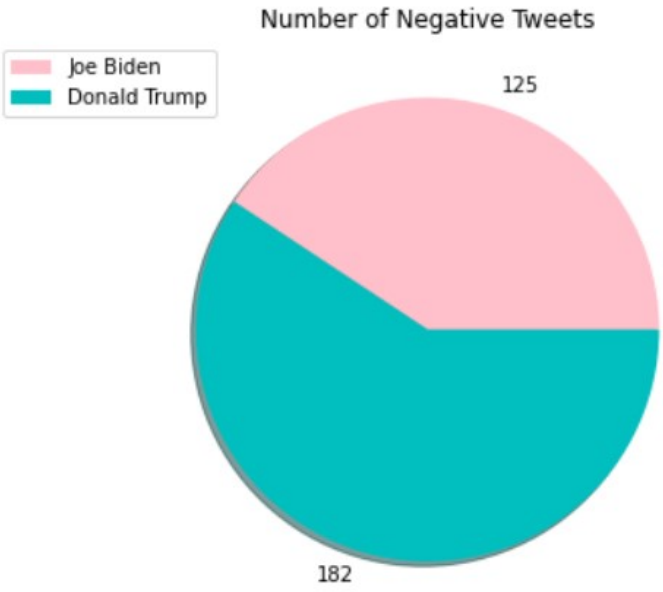

Fig 12. Negative Tweets classified based on candidates.

Election day was 11/03/2020 and from the plots below it can be seen that the frequency of tweets increased during the election day in both positive and negative cases. It can also be seen that the negative tweets increased as the election day approached but the number of positive tweets was still was more in comparison of both.

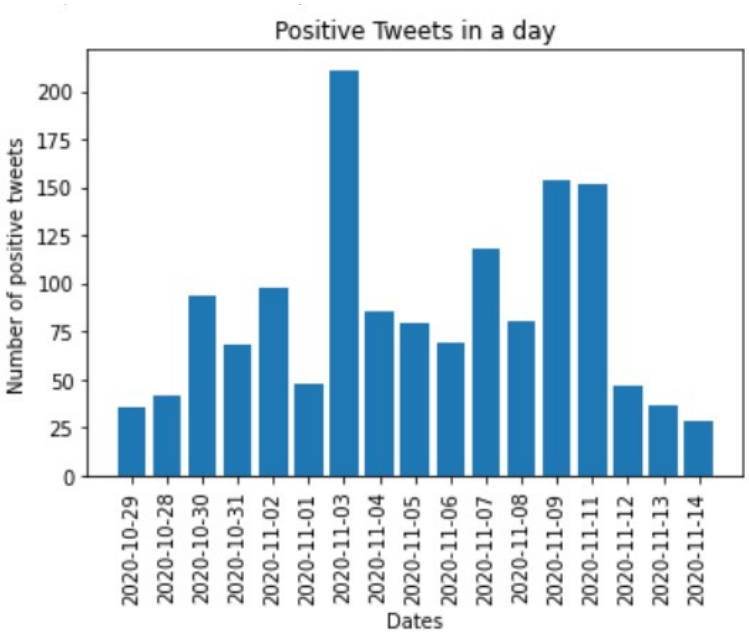

Fig 13. Positive Tweets for different dates around election day.

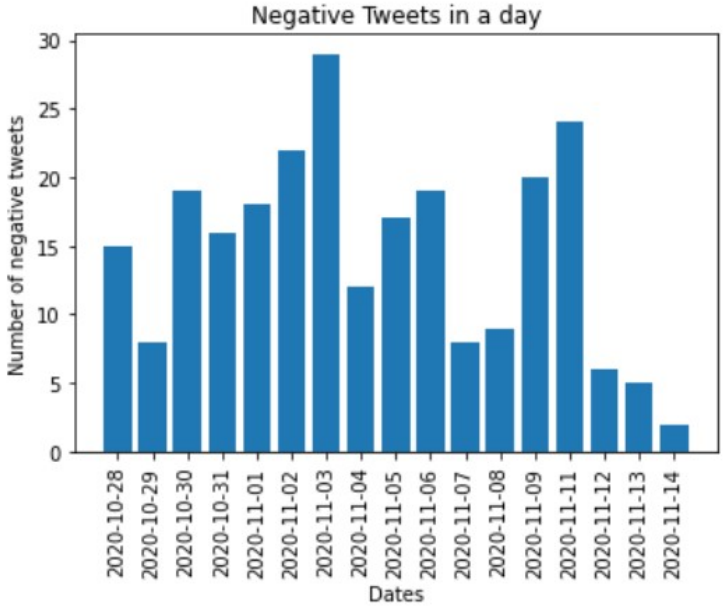

Fig 14. Negative Tweets for different dates around election day.

For the frequency of positive and negative tweets for Donald Trump it can be seen that the tweets were maximum on election day around 39 negative tweets and 55 positive on election day for this dataset and the tweets decreased to half on the other days.

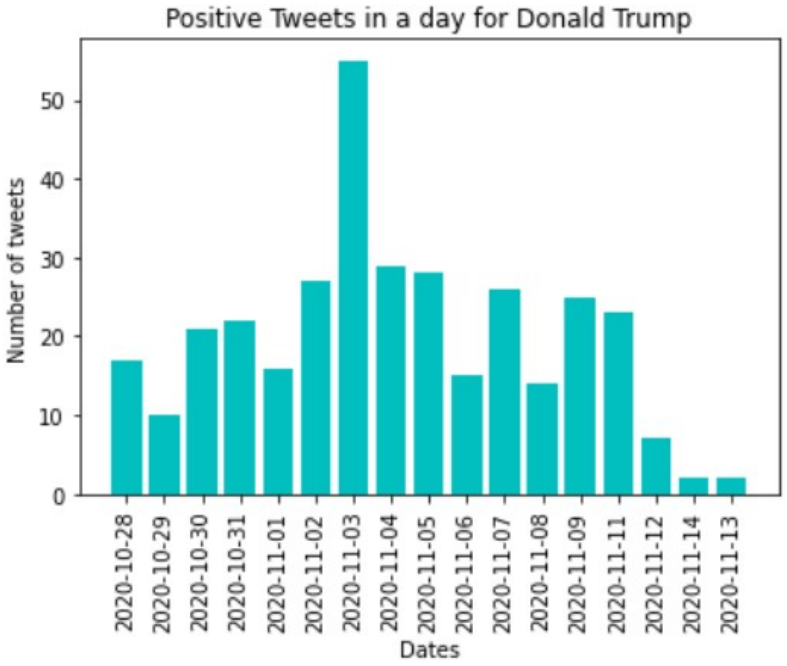

Fig 15. Number of positive Tweets for different dates for Donald Trump.

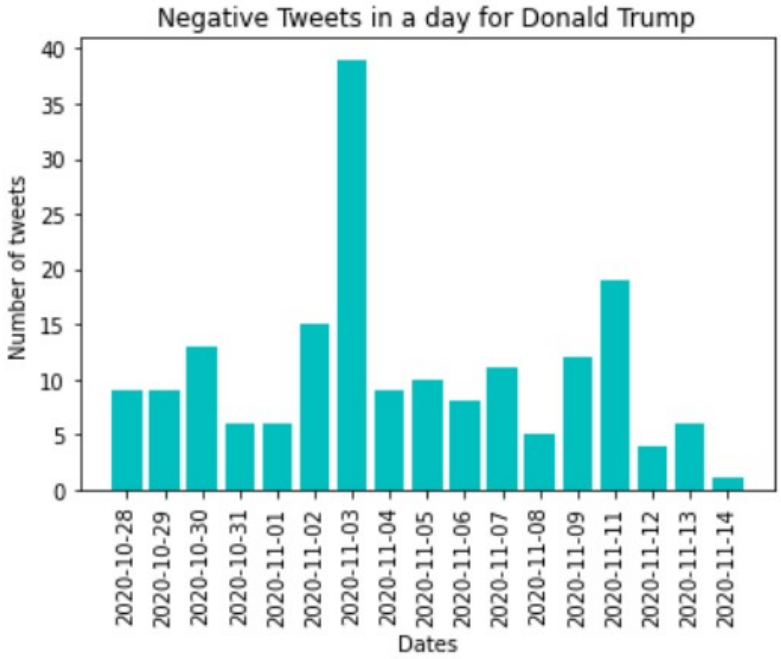

Fig 16. Number of negative Tweets for different dates for Donald Trump.

For Joe Biden the tweets. Both positive an dnegative increased on 11/09/2020 which is six days after the election day and for the other days there is not much difference or anything strange about number of tweets.

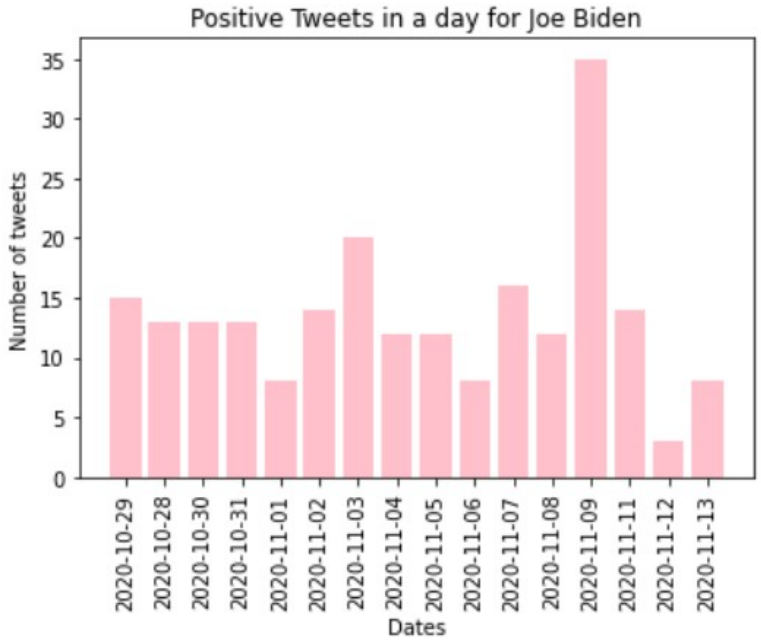

Fig 17. Number of positive Tweets for different dates for Joe Biden.

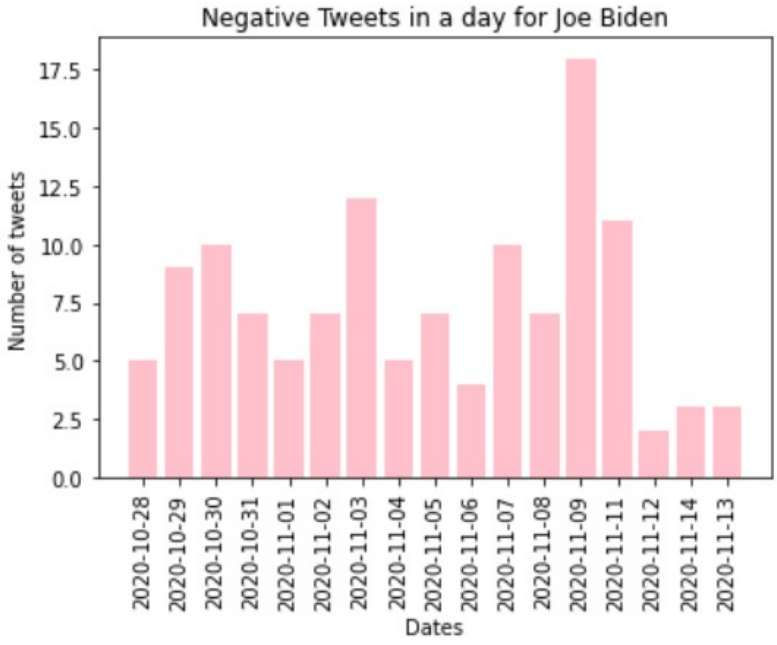

Fig 18. Number of negative Tweets for different dates for Joe Biden

Amongst the positive and negative tweet data seen it is shown in this plot that retweet from 339 positive tweets(Fig 11) for Trump 250 are retweets and from 216 positive tweets for Biden(Fig 11) 150 are retweets. From 182 negative tweets for Trump (Fig 12) around 148 are negative retweets and from 125 negative tweets for Biden(Fig 12) 100 are retweets.

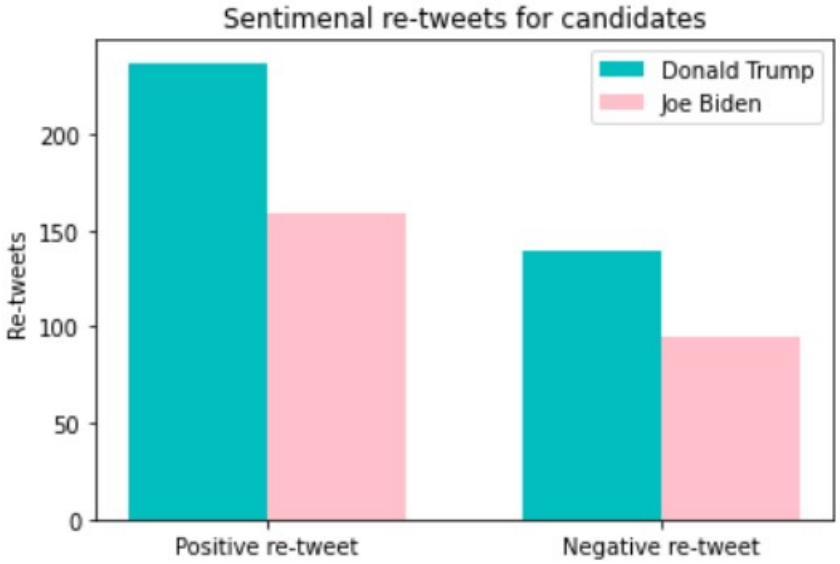

Fig 19. Retweets based on sentiments for Joe Biden and Donald Trump

The data plots for based on states is not that clean because the state column for the tweets had many garbage values. These values if deleted or changed would mess with the data integrity and if given a random value would musicality the data. Below are the plots. The positive values because they were too many the plots for that is too small. Even though the positive plots cannot be clearly seen, I was able to zoom in and the analysis is as follows. For Biden majority of negative tweets are from California. And the majority of negative tweets for Donald Trump are seen to be from USA. For positive tweets from both candidates most tweets were seen to be from USA. With Virginia being the second for majority of positive tweets for Joe Biden followed by Michigan. For trump, Florida, LA and Texas show high positive tweets.

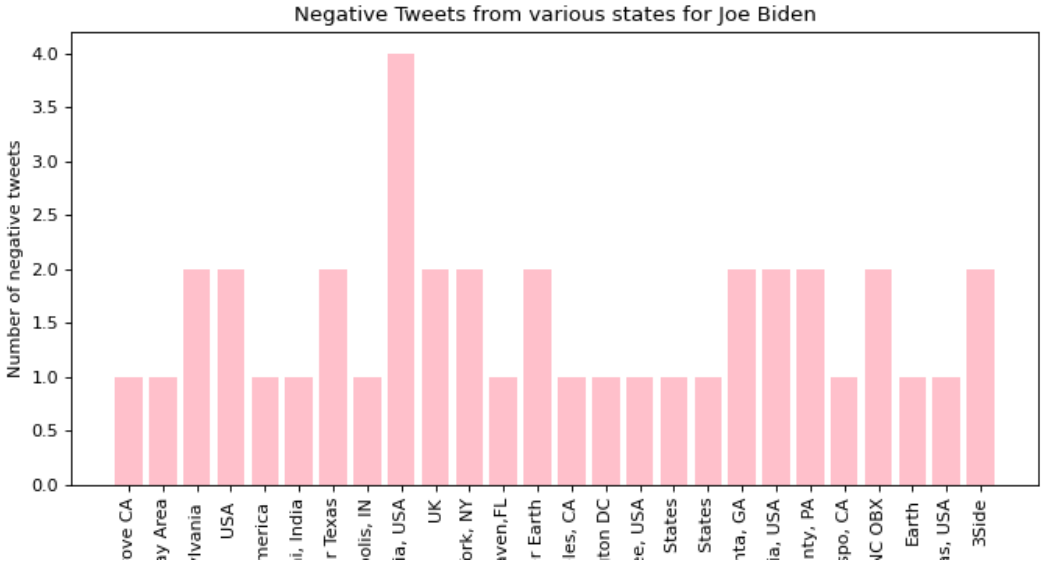

Fig 20. Negative tweets based on states for Joe Biden

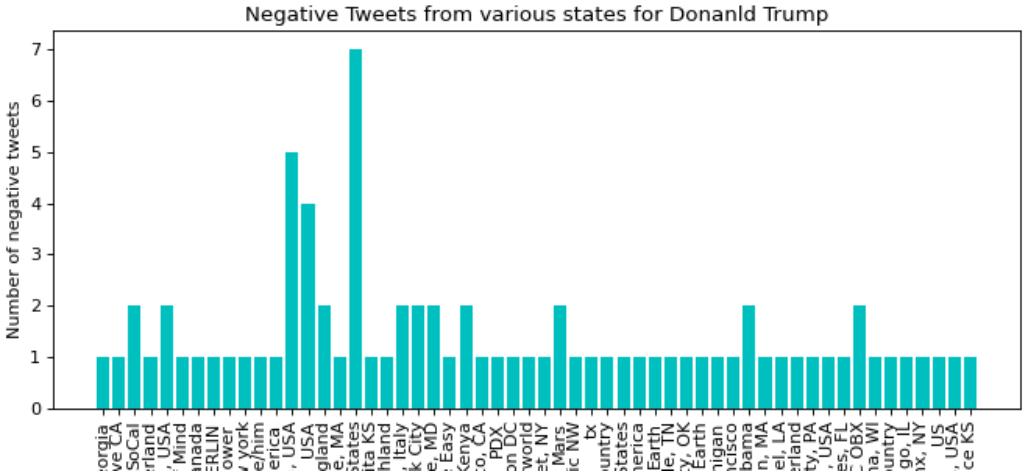

Fig 21. Negative tweets based on states for Donald Trump.

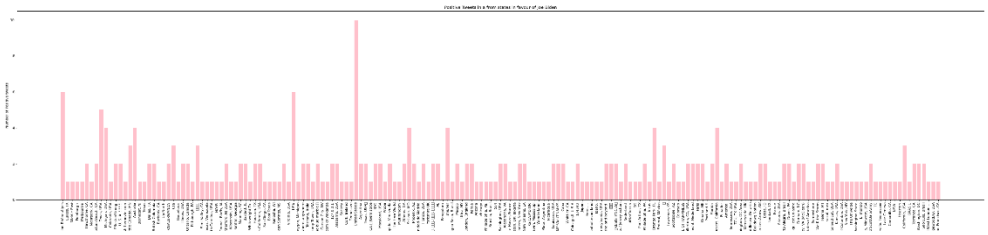

Fig 22. Positive tweets based on states for Joe Biden

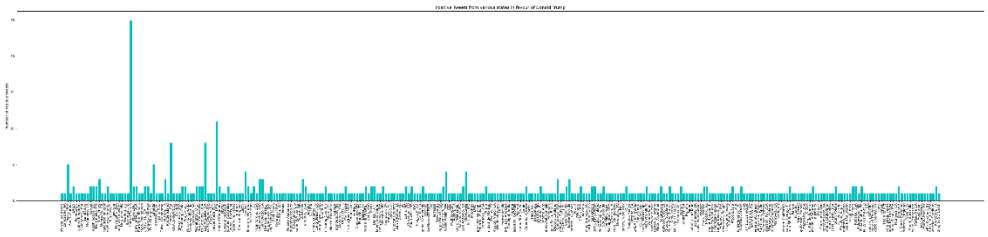

Fig 23. Positive tweets based on states for Donald Trump.