INSY 5378 Group Project 1: Social Media Analytics

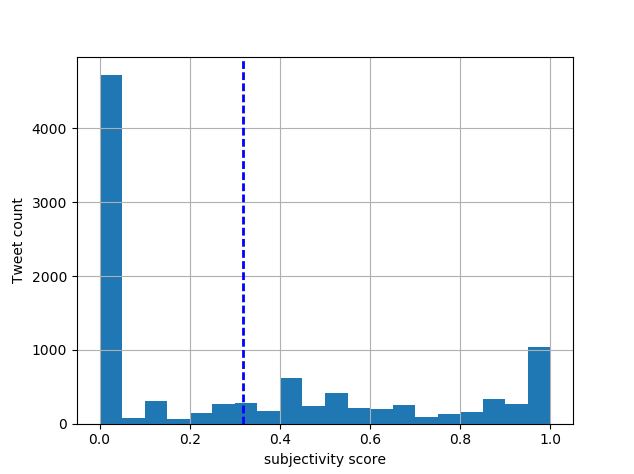
1. **Twitter Streaming API:**

We have implemented the Tweepy module and used track=KEYWORD to filter the tweets from the Twitter Streaming API. The keywords specified are ‘Trump’ and ‘Donald Trump’ and ‘trump’ . The output of the file is fed to a text file. Later this content is read from the file and we have used json model to easily scrap ‘text’ from the collected tweet corpus.

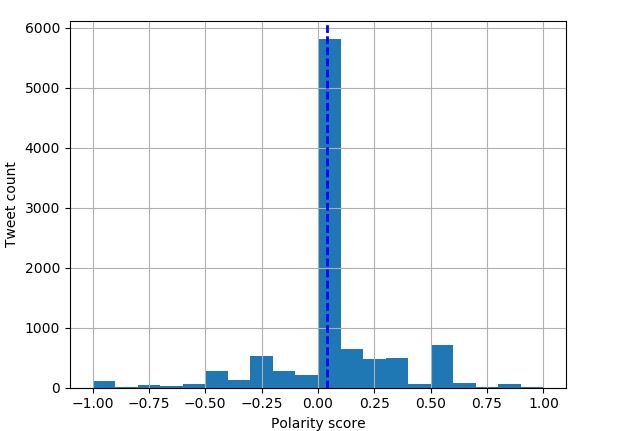
1. **Sentiment Analysis**: Positive and Negative emotions are estimated for the collected 10k tweets by estimating the values of polarity and subjectivity. TextBlob is used to conduct sentimental analysis. The aggregate values of Polarity and Subjectivity for the collected tweets are represented on Histograms and the average value for the Subjectivity and Polarity scores are calculated as follows:

|  |  |
| --- | --- |
| **Subjectivity** | **Polarity** |
| 0.31783 | 0.04157 |

**Subjectivity Histogram**

****

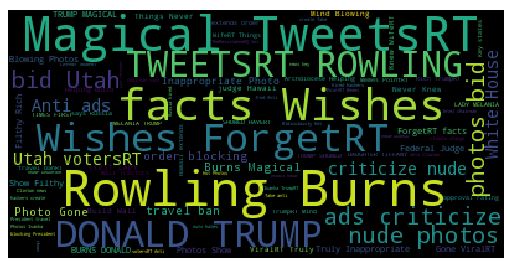
**Polarity Histogram**

****

**Note:** Vertical Dotted line on the histograms shows the average values of Subjectivity and Polarity scores.

**C. Word Cloud:**

The word cloud is generated for the 10k tweets based on Trump after preliminary cleaning and removal of stop words before feeding into the word cloud module. The image below displays the most frequent words in the tweets collected.

****

1. **Topic Modelling:**

We have conducted topic modelling on the 10K tweets collected using NMF and LDA with 3 different combinations of 20,10,5 topics with 200,100,150 passes respectively. The results of the same are shown below

**NMF:**

****

**LDA:**

****

In the attached document for LDA topic modelling, we have highlighted the models that gives the least Log Perplexity value for the collected 10 K tweets and also for the 1000 tweets collected in each of the regions(North, South, West, East and Central).

**Log Perplexity:**

|  |  |  |
| --- | --- | --- |
| Num\_Topics | Passes | 10k Trump Tweets |
| 20 | 200 | -7.456 |
| 10 | 100 | -7.678 |
| 5 | 150 | -7.839 |

Tabular values for region-wise Log perplexity values are provided in section E

1. **Geographic Variation:**

The above steps are repeated and analysis is performed for different locations across the United States. i.e., North, South, East, West and Central regions.

Below are the Longitude and Latitude coordinates for the different regions covered inside USA:

North - [-116.75,42.18,-90.91,47.77]

South - [-108.62,28.29,-84.76,36.19]

East - [-84.50,31.60,-71.49,44.76]

West - [-123.43,36.79,-113.33,48.65]

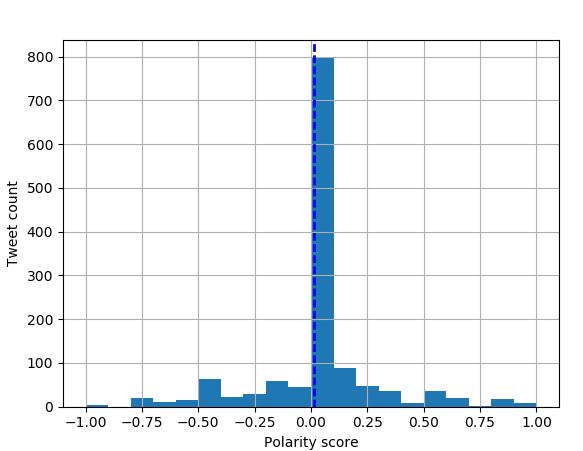
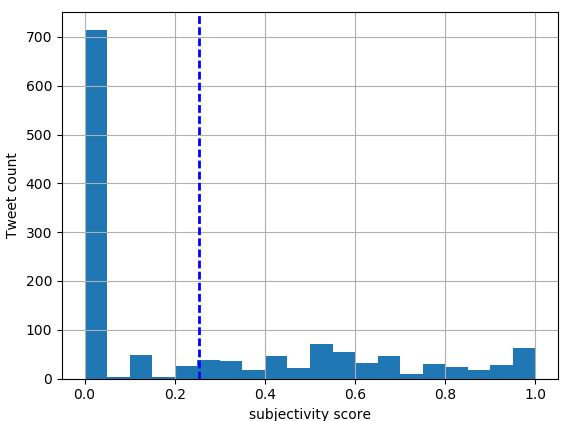
Central - [-108.54,37.02,-84.17,44.79]

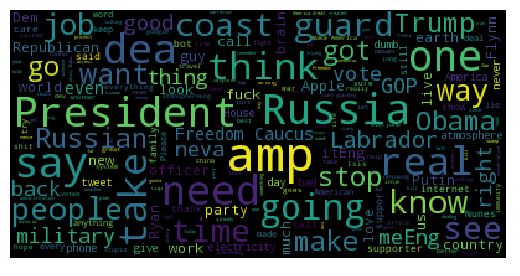
|  |  |  |
| --- | --- | --- |
| **Region** | **Subjectivity** | **Polarity** |
| North | 0.25397 | 0.01418 |
| South | 0.33061 | 0.05667 |
| East | 0.30477 | 0.02599 |
| West | 0.31029 | 0.01485 |
| Central | 0.32831 | 0.03277 |

* Looking at the Polarity average value across different regions, the tweets were mostly not positive. At the same time, the mean value did not fall on the negative side.
* Compared with each other, tweets from the South are slightly more positive than others**.**

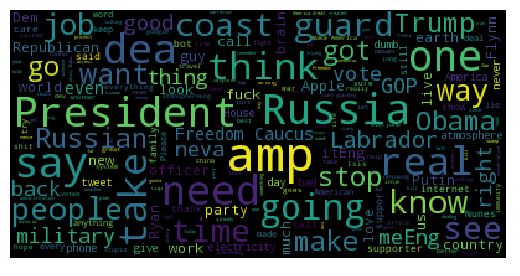
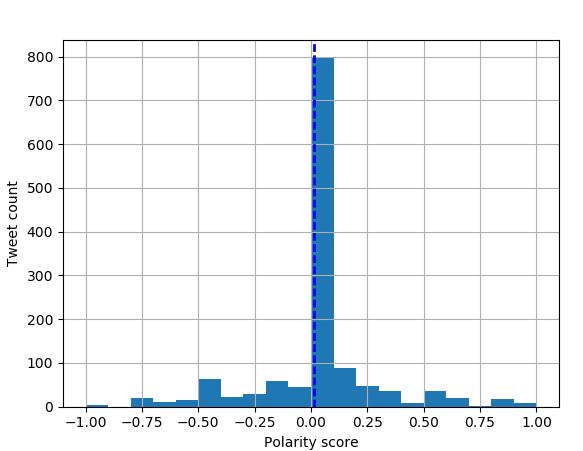
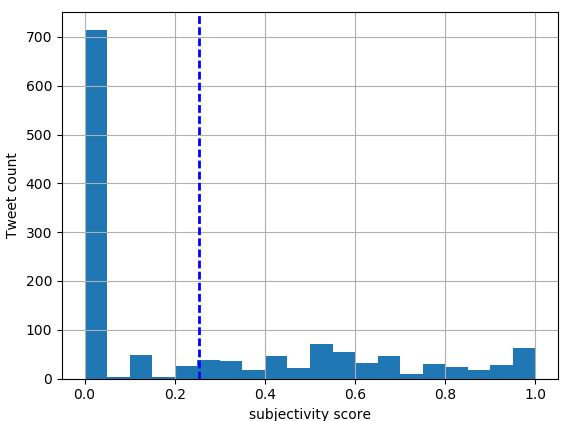
For each region, the sentiment analysis is performed and the following readings are observed. Also, Word cloud is generated for the different regions as shown in images below.

**North**

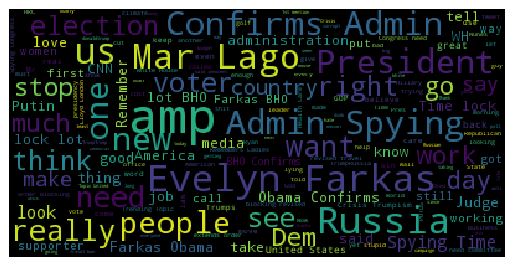
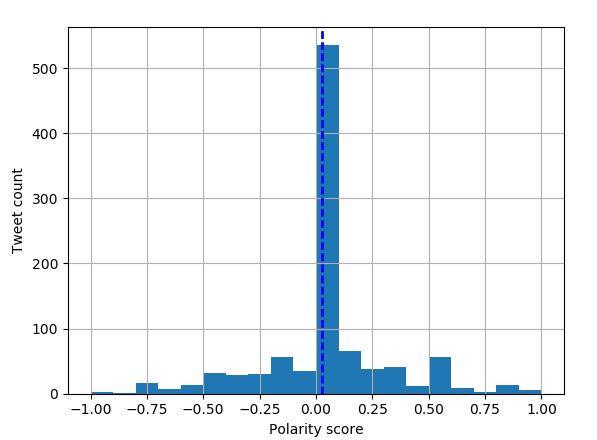
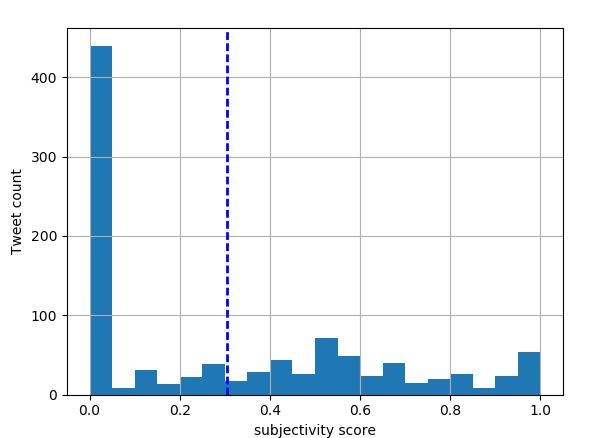
****

****

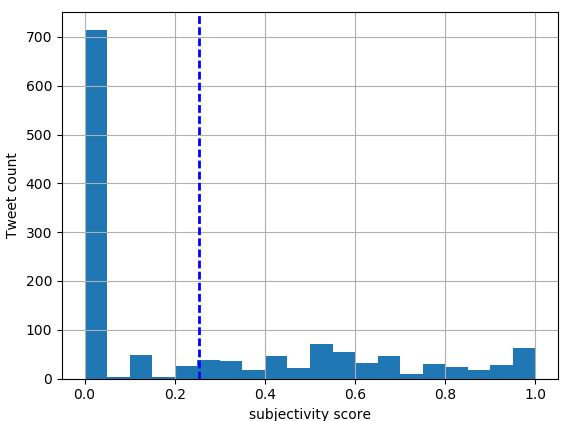
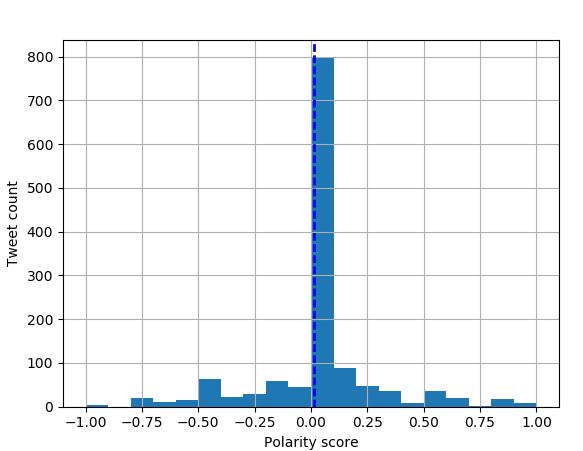
**South**

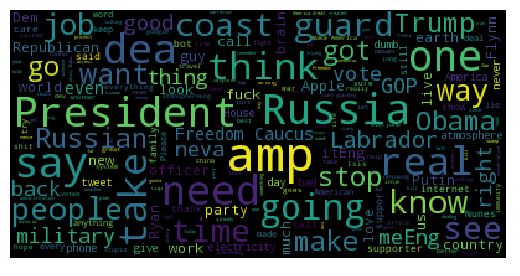
****

**East**

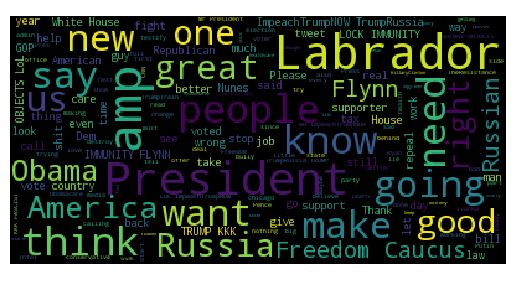
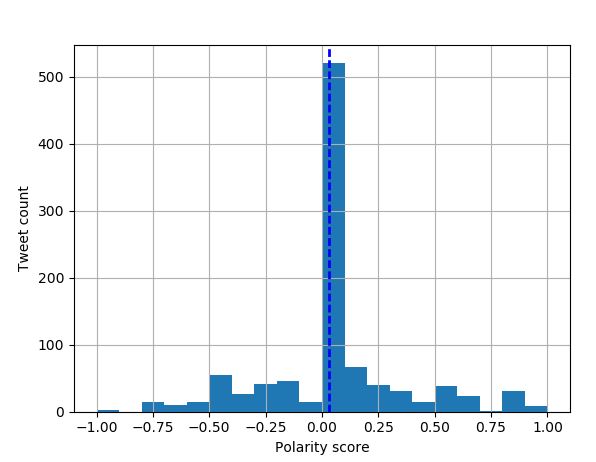
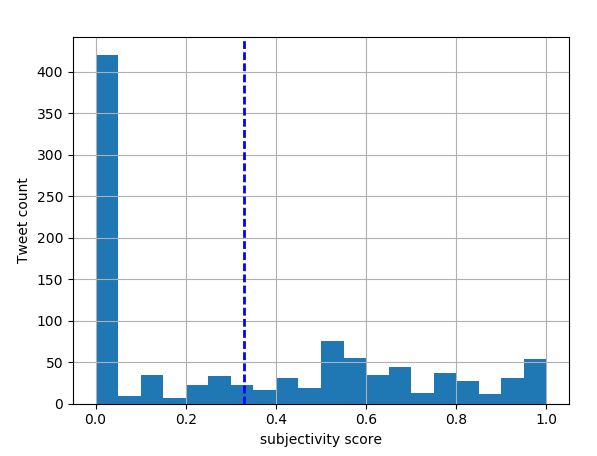
****

**West**





**Central**

****

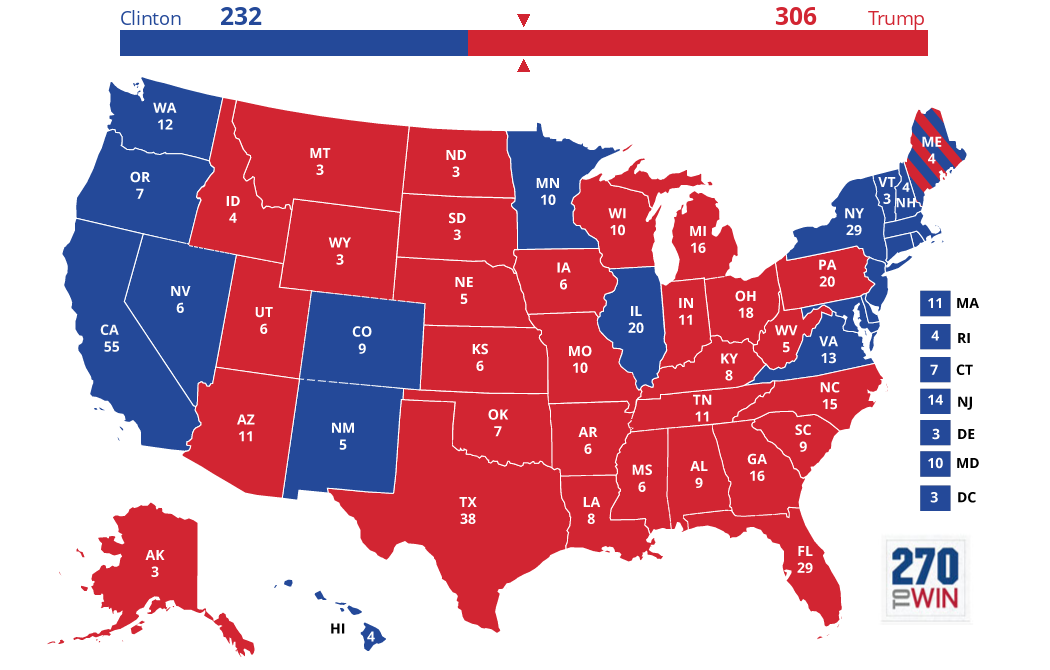
**Topic Modelling:**

Again, we have conducted topic modelling for different regions(North, South, East, West and Central) using NMF and LDA with 3 different combinations of 20,10,5 topics with 200,100,150 passes respectively. As mentioned above, the results of the same are provided in the attached documents in section D.

**Log Perplexity values**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Num\_Topics | Passes | East | West | North | South | Central |
| 20 | 200 | -9.834 | -9.07 | -9.28 | -9.307 | -9.45 |
| 10 | 100 | -9.052 | -8.177 | -8.92 | -8.98 | -9.076 |
| 5 | 150 | -9.075 | -8.773 | -8.56 | -8.63 | -8.704 |

1. **Insights:**



* As observed from section B, the average Polarity(0.056) values are observed to be the maximum thus matching the approval that Southern regions have the most support for trump.
* Also, the Bay area covering the western regions where Trump has the least support as per 2016 Election results, the polarity(0.014) scores are very low.
* It is also interesting to observe that the tweets from the Northern regions also have lesser subjectivity (0.25) and polarity(0.014) values. Though it is contradictory that Trump has the most support in these regions unlike the West, it also shows a lesser subjectivity value in the north than that of West.
* For central US, the unique topic is inferred to be ‘disgrace’ based on topic modelling.
* For South US, the unique topic is inferred to be ‘draintheSwamp’ based on topic modelling.
* For North US, the unique topic is inferred to be ‘collusion’ based on topic modelling.
* For East US, the unique topic is inferred to be ‘nightmare’ based on topic modelling.
* For West US, the unique topic is inferred to be ‘agonized’.
  + West US has got more negative words such as ‘cheated’, ‘agonized’ in topics and it accounts for our sentimental analysis which shows the least polarity score.
  + South US has maximum polarity score compared to other regions which is because it has got more positive words such as ‘saved’, ‘TrumpCare’,’freedom’ as per our topic modelling analysis
* Central and West US has got some similarities in topics such as ‘WTF’, ‘liar’, ’idiot’ and we think the reason for similarity could be from a common topic ‘lies’.

**Overall conclusion:**

From our analysis, we strongly think that people are not happy about Trump being in the power. The polarity score is less than 0.1 and that is the reason we feel sentiments are not positive.

**References:**

<https://www.quora.com/How-do-I-remove-stopwords-from-a-file-using-python>

<http://stackoverflow.com>

<http://adilmoujahid.com/posts/2014/07/twitter-analytics/>