INSY 5378 : Data Science

Group Project -1

Social Media Analytics

by

Sushidhar Jayaraman (#1001400523),

Prashanti Chandrasekaran (#1001),

Pradeep Kumar Madhangopal (#1001448700)

A. **Data Collection :** Using Twitter Streaming API, collect 1M tweets including the keyword "trump". You may use tython for tweet collection. As shown in the lecture, you can use `track=KEYWORD` to get keyword-filtered tweets. (Note: In case you lose the API connection with Twitter, you may need to run your code multiple times to collect sufficient number of tweets.)

- Obtained CONSUMER_KEY, CONSUMER_SECRET, ACCESS_TOKEN, ACCESS_TOKEN_SECRET from Tweeter API.
- Included Keyword = 'trump' in twitter streaming code.
- Using the above keys in Streaming code, we retrieved 1M tweets from twitter API and saved the output as a dictionary in .json file format.

Data Pre processing :

The real world data in general contains missing values, discrepancies, noise, unwanted data, etc. To overcome these issues, we need to follow certain data pre processing procedures.

Given below are some of the data preprocessing methods that we followed in our project:

Content :	**Action performed :**	**Reason :**
@ symbol	Removed	The @ symbol is generally used to address a specific user in response to a comment made in that context. Having this in our analysis would not yield any meaning to the data.
# (hashtaged) words	Removed	The hashtag '#' symbol is widely used in tweets. The word content following this symbol is generally a combination of a few words, which doesn't mean anything in an appended form. Hence we removed these words.
URLs (http), Weblinks (www), etc	Removed	These are part of unwanted data and hence it has been removed in our analysis.
Upper case characters	Converted	We converted all the Uppercase characters to Lower case.
&, ~ and other symbol	Removed	These characters are of no importance in our analysis and hence has been removed.
Punctuation (!, ?, ', ", . , etc)	Removed	These are part of unwanted data and hence it has been removed in our analysis.
Numbers	Removed	We do not require Numerical data for our analysis. Hence we removed all the numerical digits.
Stop words	Removed	Stop words are to be removed in data preprocessing as they individually don't contribute any meaning to text mining. Apart from prepositions, articles, etc., we in our project manually included a few more stop words like : 'RT', 'Trump', 'Donald', 'Donald Trump', 'Trumps'.

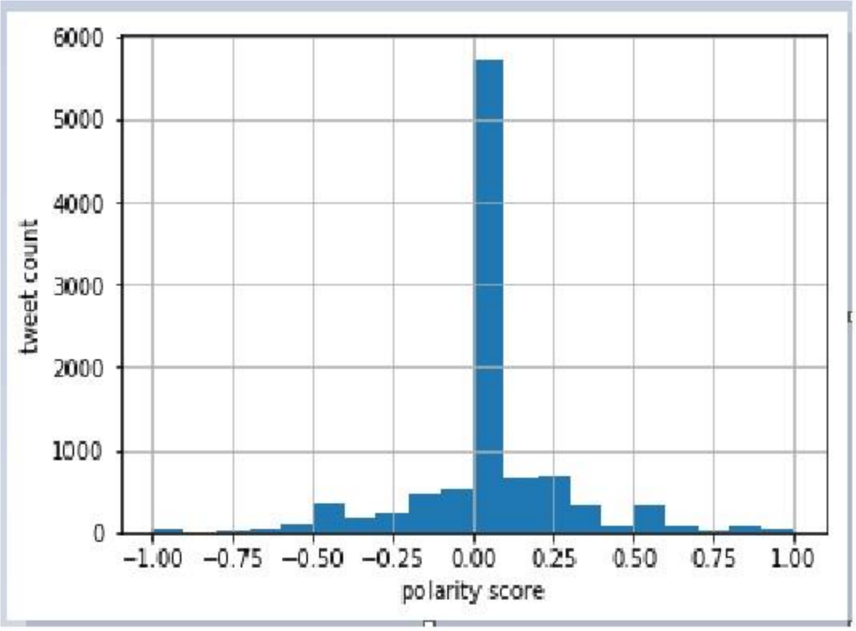
B. **Sentiment Analysis :** Using TextBlob, calculate the polarity and subjectivity scores for each tweet in the 10K tweet corpus. Summarize the calculated scores with histograms using Matplotlib, where X-axis is the score and Y-axis is the tweet count in the score bin. Also, provide the average of the polarity and subjectivity scores.

B.1) Sentimental Analysis for 1M tweets :

B.1.1) Polarity :

Polarity in sentimental analysis refers to the direction towards which the data relates to. The expressed opinion or the content of analysis, could be of a Positive polarity or a Negative polarity or a Neutral polarity.

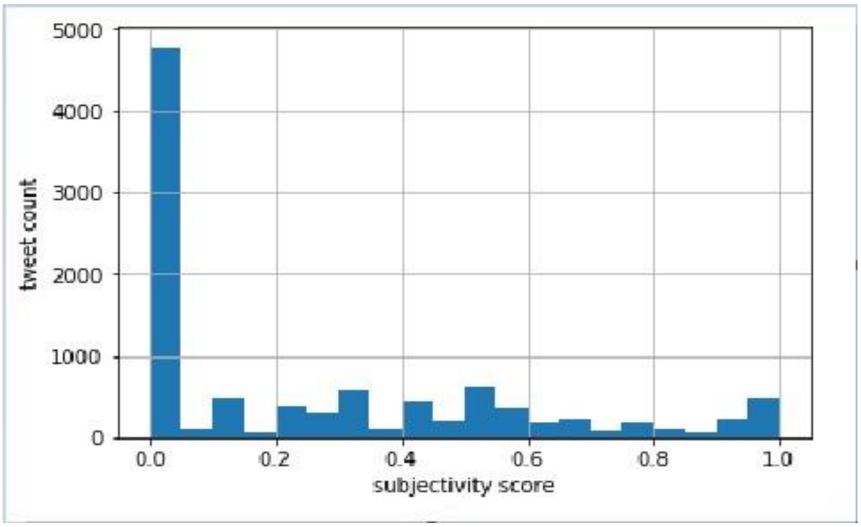
Average polarity: 0.0275423988928



B.1.2) Subjectivity :

Subjectivity of data simply refers to how related the datas are with each other. Whether they pertaining to the main context are not.

Average Subjectivity: 0.258172784287

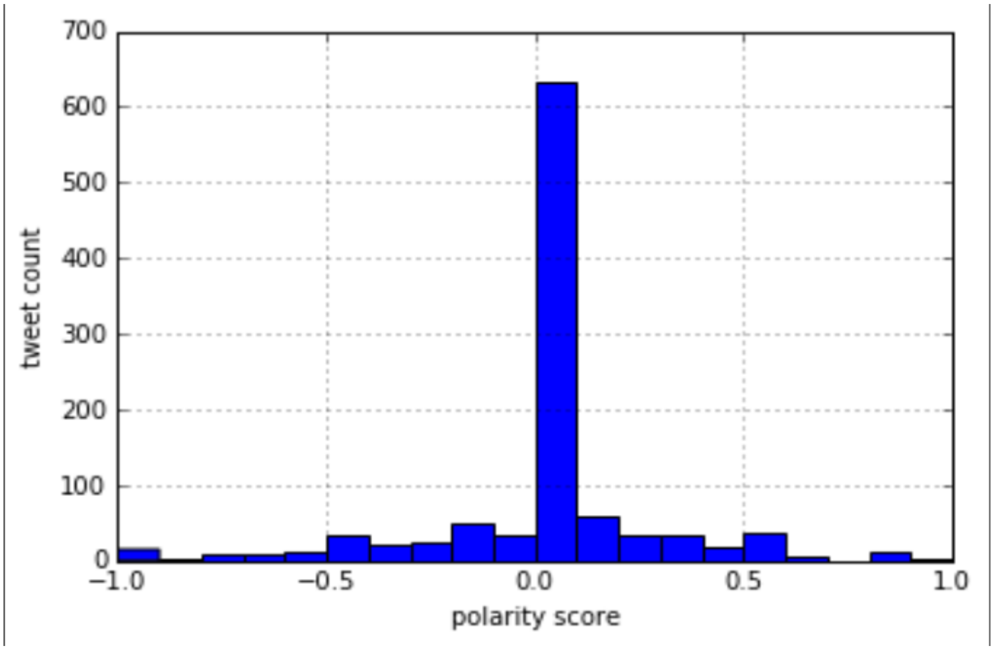


B.2) Sentimental Analysis for five regions - Arizona , California , Florida , NewYork & Texas:

B.2.1) Arizona Region :

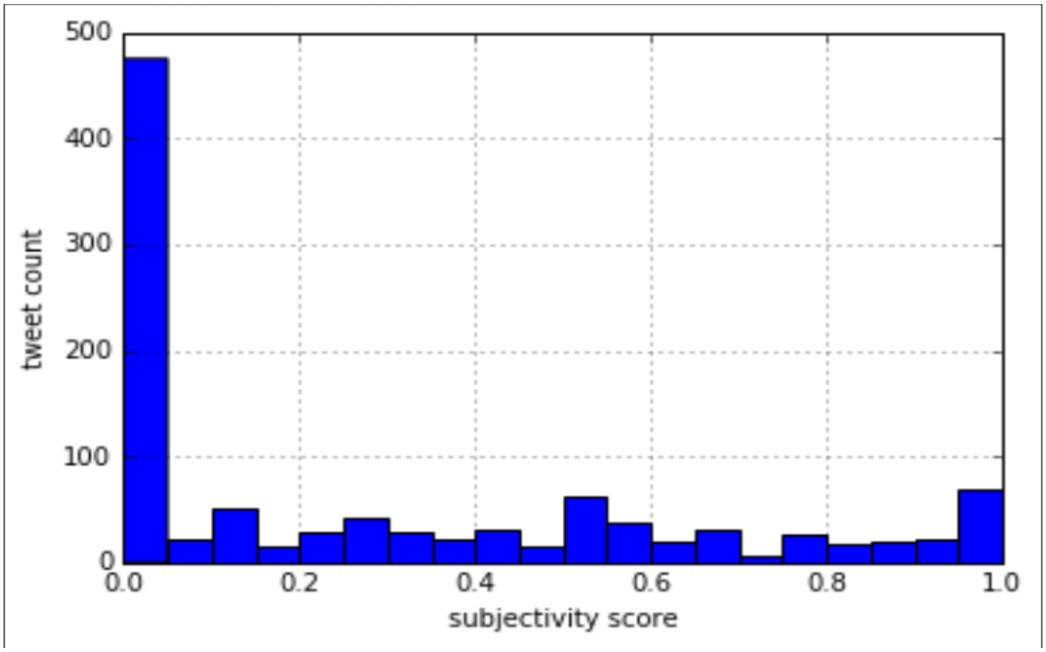
B.2.1.1) Arizona Region Polarity :

Average polarity: 0.00750423972048



B.2.1.2) Arizona Region Subjectivity :

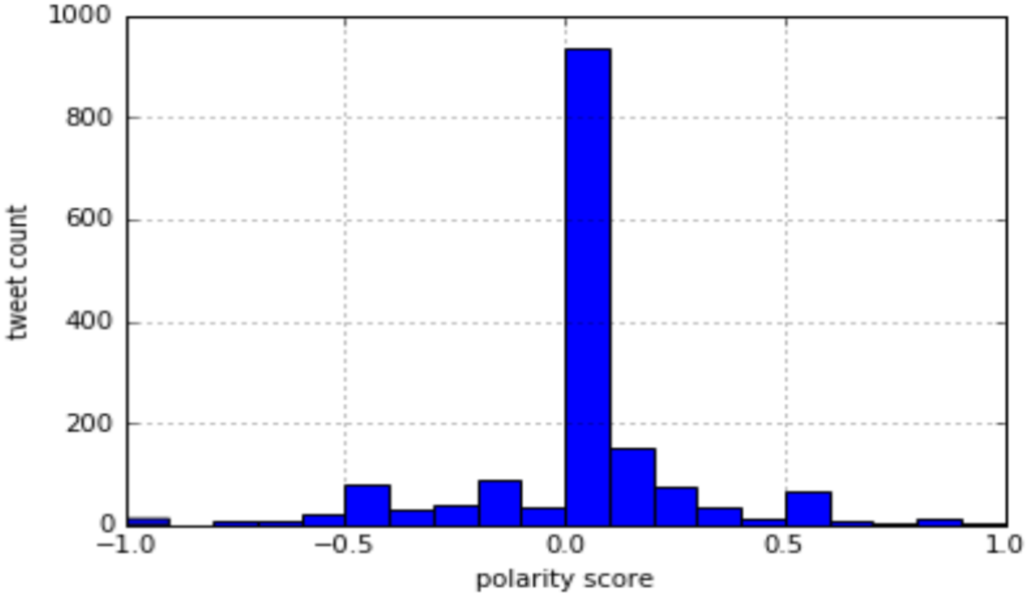
Average subjectivity: 0.282924884068



B.2.2) California Region :

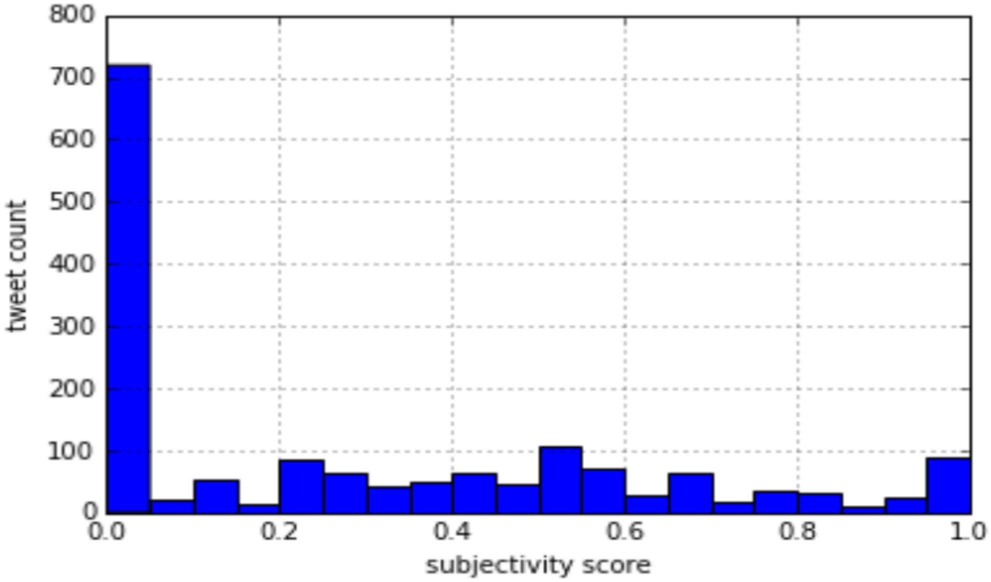
B.2.2.1) California Region Polarity :

Average polarity: 0.0150134961649



B.2.2.2) California Region Subjectivity :

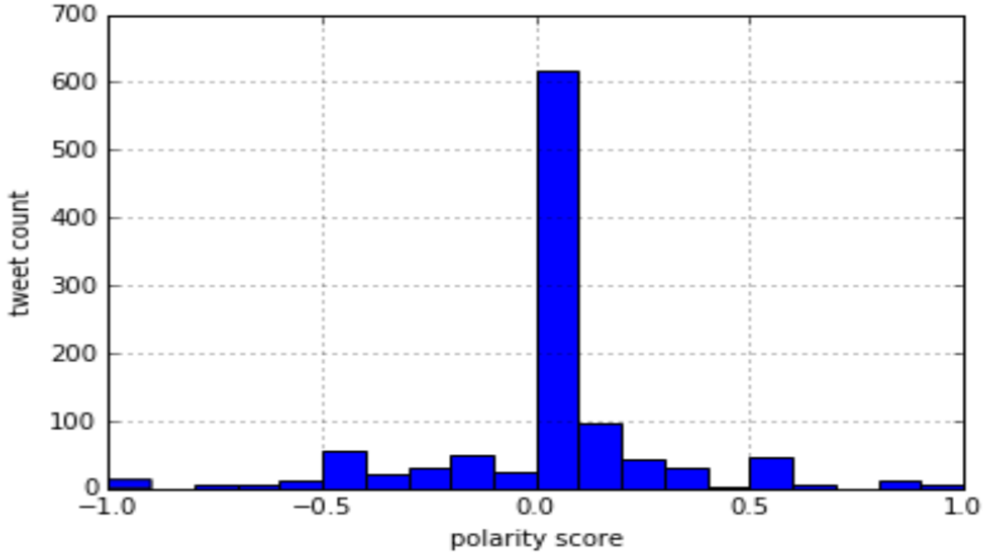
Average subjectivity: 0.285502506074



B.2.3) Florida Region :

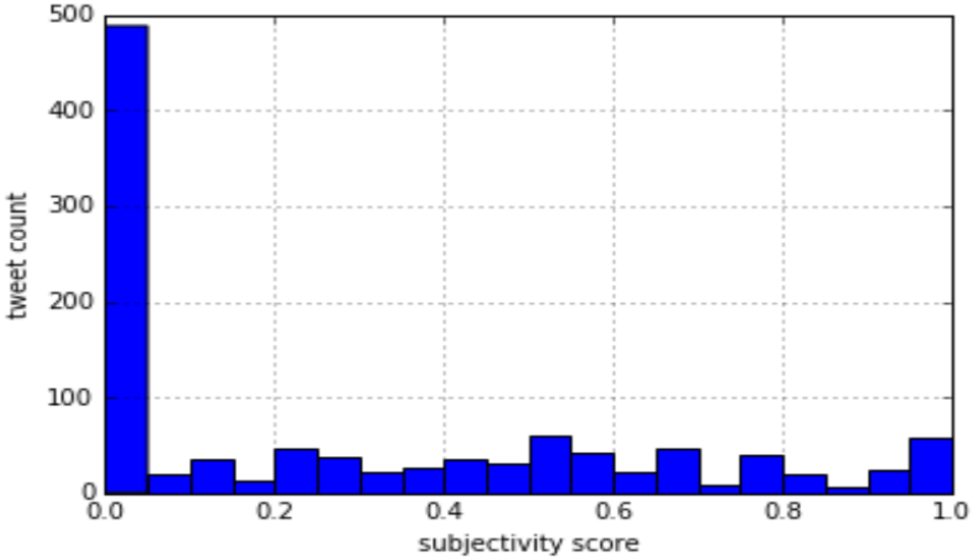
B.2.3.1) Florida Region Polarity :

Average polarity: 0.0110254245652



B.2.3.2) Florida Region Subjectivity :

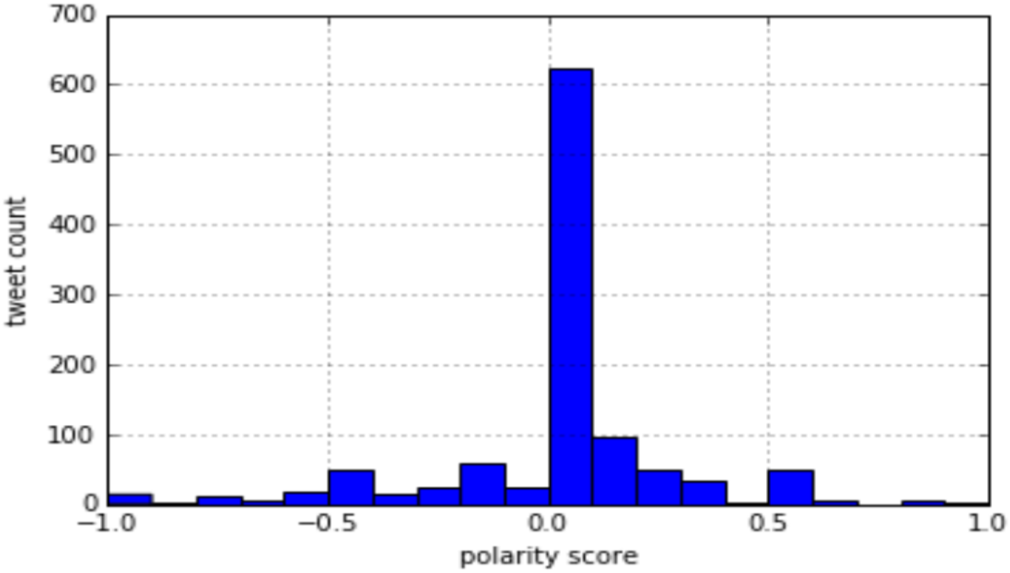
Average subjectivity: 0.287478962726



B.2.4) New York Region :

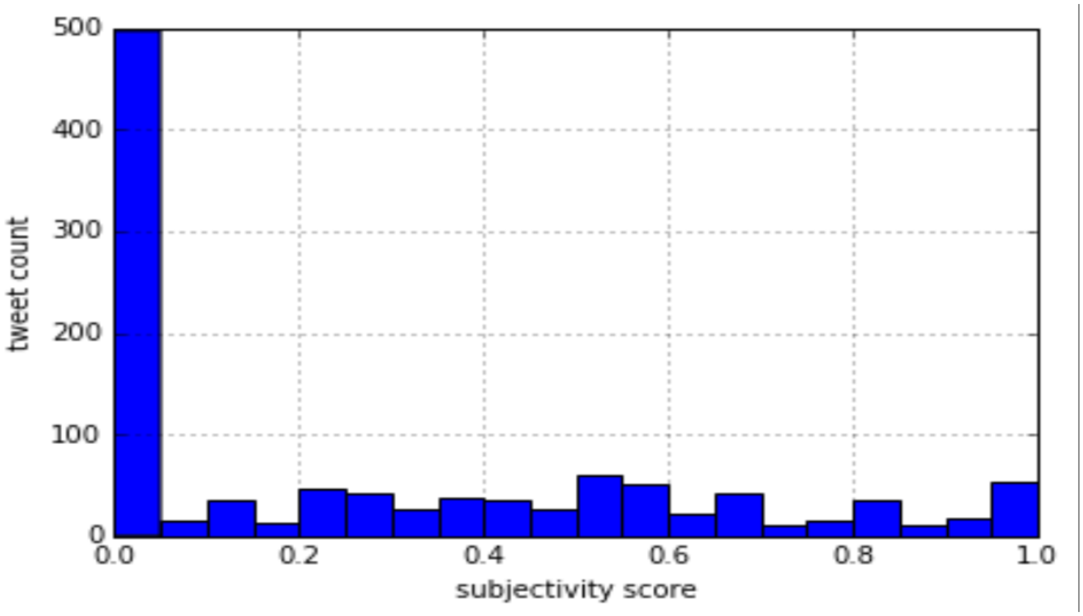
B.2.4.1) New York Region Polarity :

Average polarity: 0.00771111397997



B.2.4.2) New York Region Subjectivity :

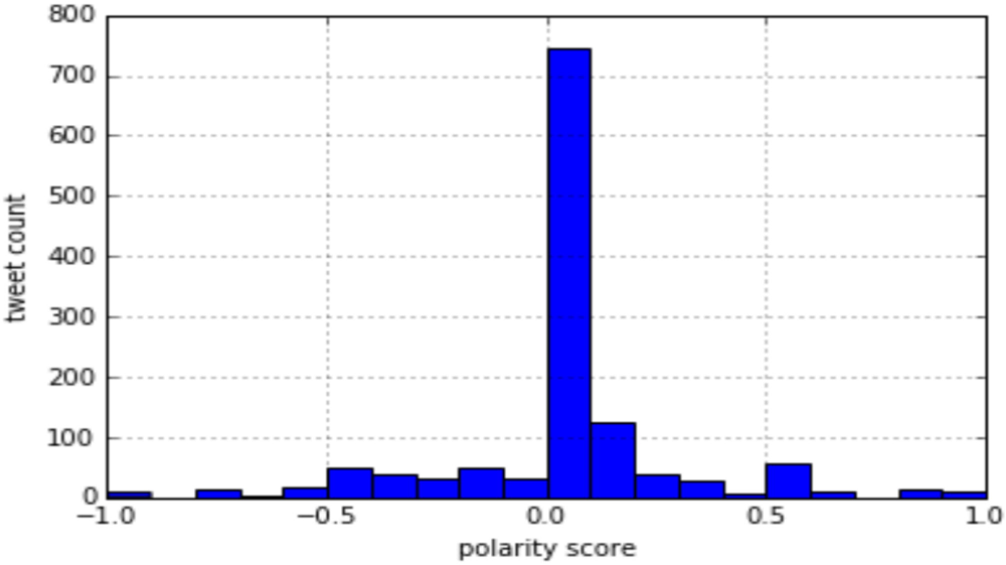
Average subjectivity: 0.280239950736



B.2.5) Texas Region :

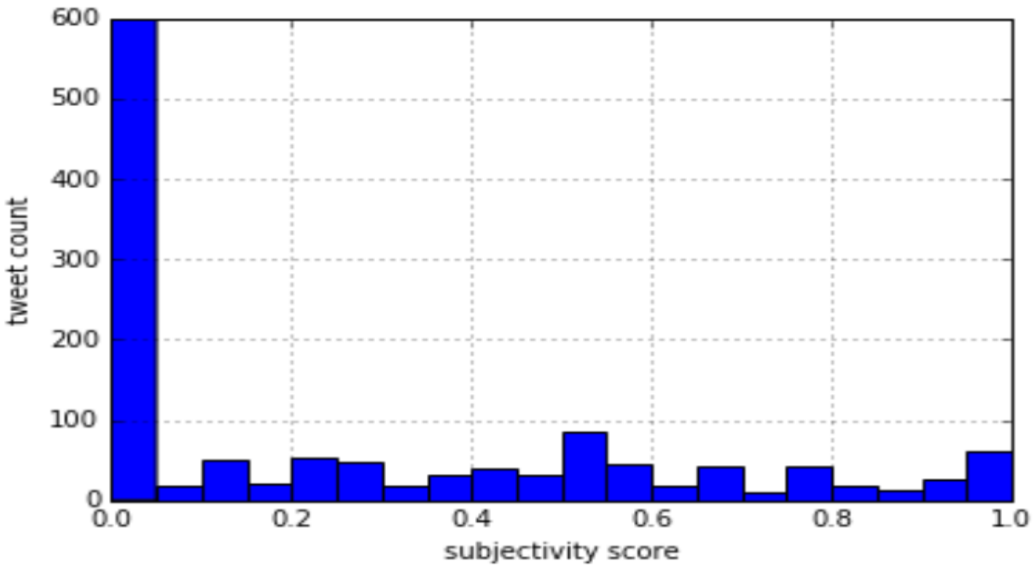
B.2.5.1) Texas Region Polarity :

Average polarity: 0.0211544562847



B.2.5.2) Texas Region Subjectivity :

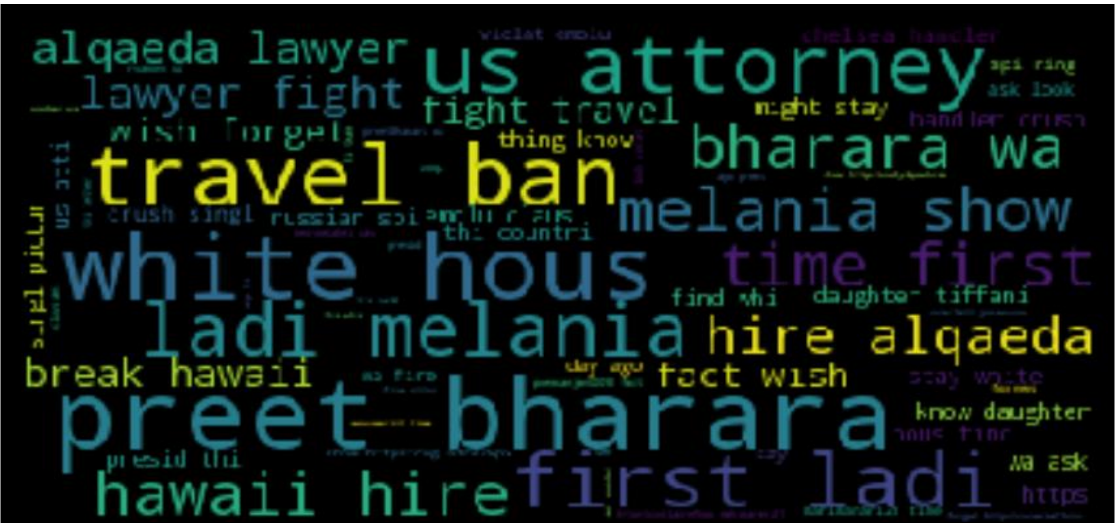
Average subjectivity: 0.270304768884



C. **Word Cloud :** Create a word cloud from the collected 1M tweets. Please remove stop words and do stemming before feeding into the word cloud module.

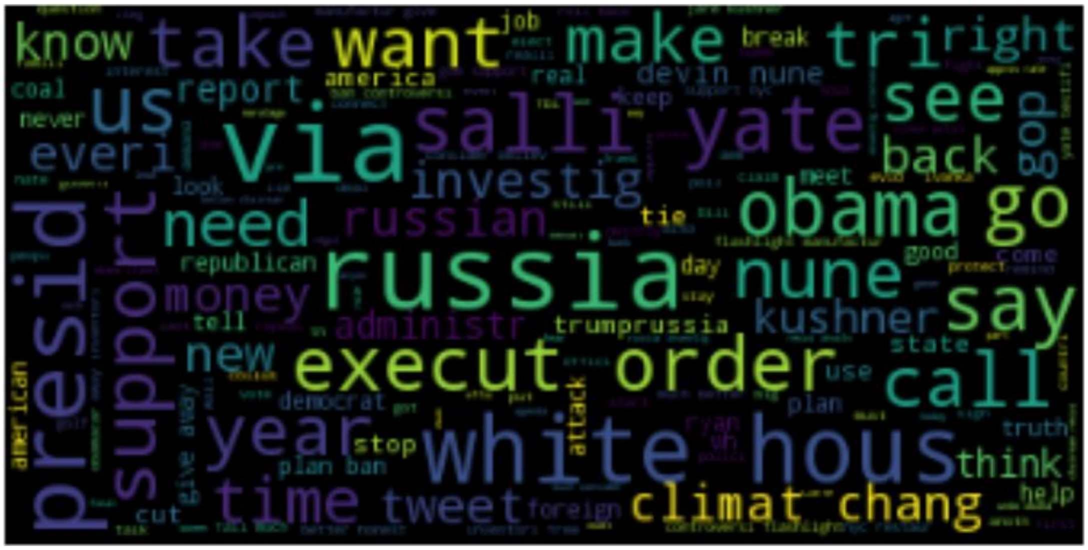
C.1) Word Cloud for 1M tweets :

z

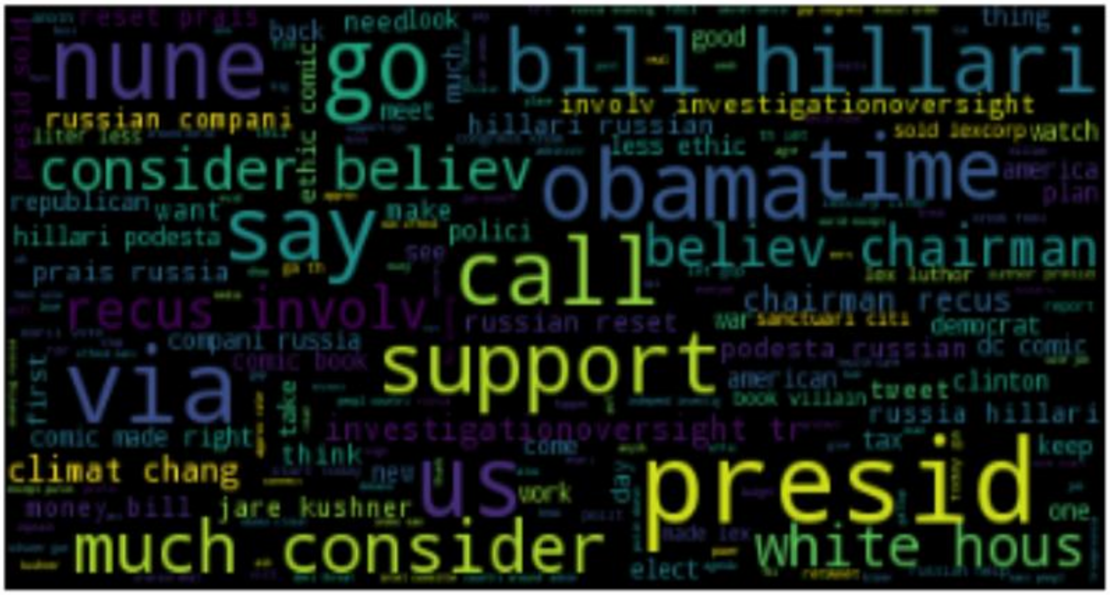


C.2) Word Cloud for five regions :

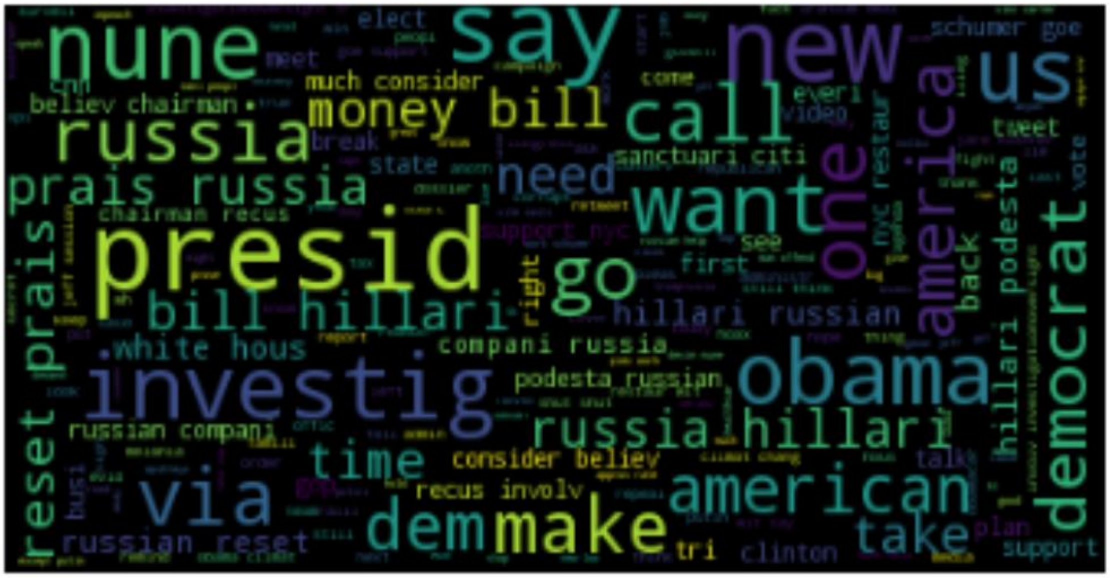
C.2.1) Arizona Word Cloud :



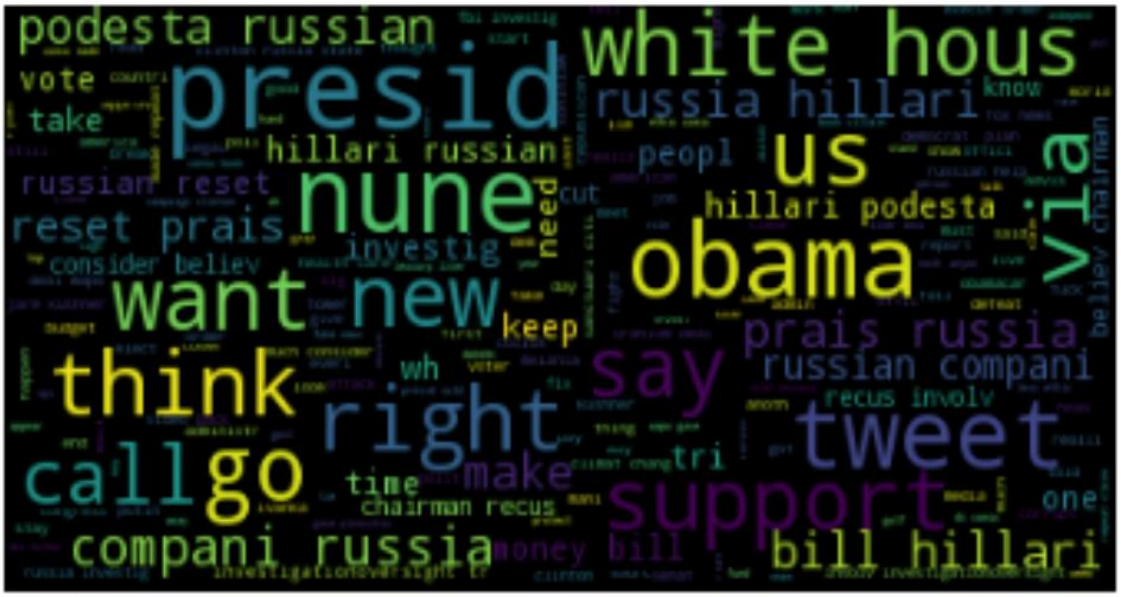
C.2.2) California Word Cloud :



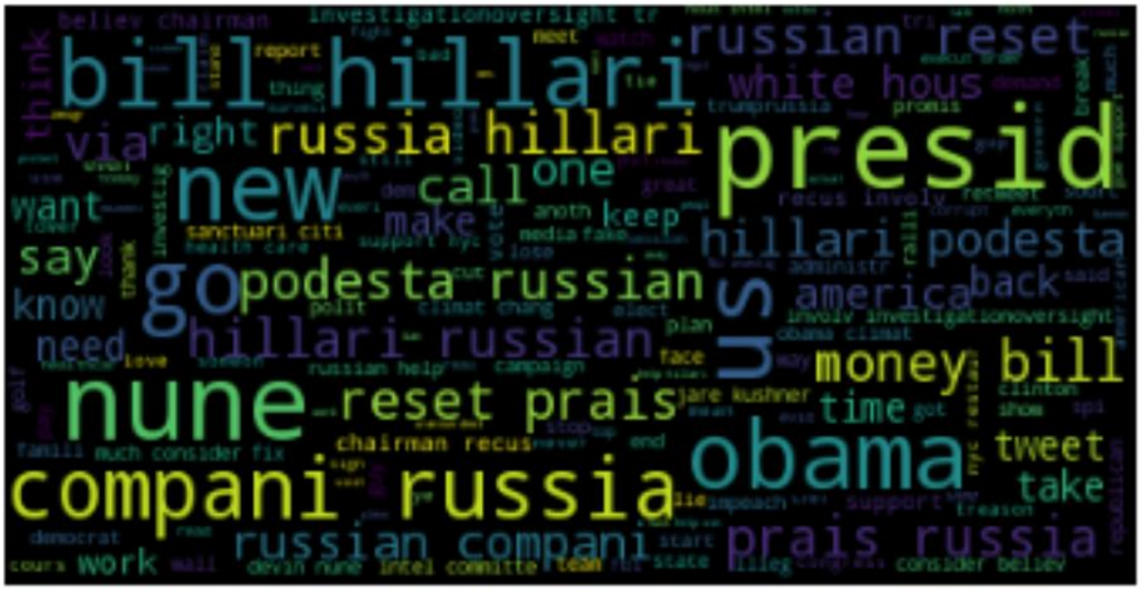
C.2.3) Florida Word Cloud :



C.2.4) New York Word Cloud :



C.2.5) Texas Word Cloud :



D. **Topic Modeling :** Conduct topic analyses on the 1M tweet corpus. You can try Non-negative Matrix Factorization (NMF) from Scikit-Learn and Latent Dirichlet Allocation (LDA) from GENSIM.

  a. Again, use appropriate stemming and remove stop words.

  b. Please vary the topic counts to find the best topic models. You only need to report the best topic model you constructed.

D.1.1) 15 topics 500 passes 5 words LDA :

```
[(0, u'0.037*"facts" + 0.033*"forget" + 0.031*"wishes" + 0.031*"we\u2019d" + 0.015*"trumps"'),
(1, u'0.053*"first" + 0.045*"white" + 0.045*"house" + 0.040*"times" + 0.033*"showed"'),
(2, u'0.040*"bharara" + 0.029*"preet" + 0.027*"trumps" + 0.024*"look" + 0.023*"asked"'),
(3, u'0.029*"ivanka" + 0.020*"evidence" + 0.017*"snl" + 0.015*"scarlett" + 0.015*"johansson"'),
(4, u'0.016*"already" + 0.016*"trumps" + 0.015*"law" + 0.014*"need" + 0.014*"supporters"'),
(5, u'0.052*"ban" + 0.051*"travel" + 0.039*"breaking" + 0.037*"fight" + 0.036*"lawyer"'),
(6, u'0.034*"bharara" + 0.030*"us" + 0.028*"preet" + 0.025*"stay" + 0.018*"attorney"'),
(7, u'0.051*"president" + 0.038*"russian" + 0.037*"get" + 0.036*"ring" + 0.034*"oh"'),
(8, u'0.022*"will" + 0.017*"just" + 0.016*"trumps" + 0.016*"never" + 0.015*"must"'),
(9, u'0.045*"trumps" + 0.033*"know" + 0.029*"daughter" + 0.028*"things" + 0.025*"tiffany"'),
(10, u'0.018*"call" + 0.017*"letter" + 0.016*"tried" + 0.015*"received" + 0.015*"bharara"'),
(11, u'0.057*"melania" + 0.013*"ivanka" + 0.011*"congress" + 0.011*"think" + 0.011*"china"'),
(12, u'0.033*"us" + 0.024*"fired" + 0.021*"just" + 0.018*"admin" + 0.016*"got"'),
(13, u'0.048*"lady" + 0.030*"picture" + 0.028*"single" + 0.027*"crushes" + 0.026*"handler"'),
(14, u'0.013*"anything" + 0.013*"thread" + 0.012*"flynn" + 0.012*"trumps" + 0.011*"foreign"')]
```

D.1.2) NMF 10 topics 10 words :

```
Topic 0: showed, lady, melania, times, care, hypocri, golfing, weeks, visited, golf
Topic 1: alqaedas, hires, lawyer, hawaii, fight, breaking, travel, ban, reed, winnable
Topic 2: wishes, facts, forget, eliminate, clocks, forward, turn, presidency, hour, correct
Topic 3: tiffany, things, daughter, know, manager, literally, figuratively, killer, tell, folks
Topic 4: white, house, stay, intruder, evidence, committee, service, secret, agent, propaganda
Topic 5: country, president, stand, retweet, new, tourism, sparring, america, construction, wins
Topic 6: chelsea, handler, crushes, single, picture, tweet, releases, navy, uss, concept
Topic 7: preet, bharara, asked, days, violations, look, fired, ago, wonder, clause
Topic 8: oh, ring, spy, want, fi, russian, took, team, busted, click
Topic 9: overseeing, investigation, major, admin, atty, got, fired, need, clinton, seeps
Topic 10: thursday, tried, letter, wednesday, received, bharara, tod, investigate, know, bharar
Topic 11: female, men, murd, homicide, victims, half, need, partner, yall, killed
Topic 12: administration, false, true, spied, president, stand, day, claims, scheduled, remain
Topic 13: memes, point, increasingly, reminder, trumpworld, chance, ethical, helping, stein, facts
Topic 14: constitution, china, bribing, mexico, prevent, potus, clause, emoluments, trademarks, blatant
Topic 15: huckabee, sanders, rising, orbit, star, sarah, onair, faced, aggressive, deal
Topic 16: snl, ivanka, scarlett, johansson, perfume, ad, complicit, day, independence, botches
Topic 17: china, money, called, thinks, florida, casinos, press, laundromats, realitythey, giant
Topic 18: experts, revised, denounced, ban, travel, read, todayng, policy, foreign, nigerianewsdesk
Topic 19: russia, obama, russian, ap, interview, spokesman, talks, months, son, oligarch
```

The features extracted are next passed on to SVM classifier.
The model built is used to predict the sentiment of the new tweets.

A baseline model by taking the unigrams, bigrams and trigrams and compare it with the feature-based model for both the sub-tasks

Results:

The feature-based model provided the improved accuracy of 77% compared to the baseline model with 62% accuracy.