Submission report : CSC 501

Assignment 4 : Team D

Team members :

Sushil Paneru V00936990
Rajneesh Gulati V00949939

This report contains the insights, design choices, data visualizations, challenges and adherence details to the rubric provided for this assignment.

Data Models and Design Choices

1. Data Preprocessing

a. **Text cleaning**

Stop words and punctuation removal - Stop words are insignificant and increases the volume of text which adds noise to the dataset. Hence, they are removed using NLTK library.

URL removal - Urls are not very informative to text analysis, especially when they are shortened (tinyurls). Hence, they are removed using regular expression to match the pattern (http\s+.*) , which is valid for all kinds of urls.

Lowercasing words - All words are converted to lowercase so that different cased words are treated the same. Eg: Twitter and twitter are treated to be the same.

Lemmatization - Lemmatization of the words was done to reduce the volume of the dataset and still preserve the context. NLTK's stemming library was avoided because we were getting irrelevant results. For eg: president was getting reduced to an invalid word presid; this made difficult to analyze the word2vec and we observed similarity score between similar words decreased with stemming applied.

Parallelization - We divided the task of cleaning text into 8 processes with the help of **dask**. It is recommended to use as many processes as logical cores available; and due to lack of time, we were not able to analyze the parallelization result with different number for processes. We were able to perform text cleaning on whole dataset (13 csv files) in **2-3 minutes**.

b. **Index addition and datetime columns**

We added index on tweet_id column for faster retrieval and also made publish_date column as datetime type. Also for Sentiment Analysis extracted just the year.

2. Data Models

a. **Trie**

We are interested in query such as find all the tweets that mention the word trump and impeachment. Thus, we need a dictionary data structure that allows us to retrieve words as key, perform union/intersection operations on value and occupy minimal space. Hashtable is a viable option but trie can efficiently share memory when common prefixed words are stored which is why trie was chosen over hashtable.

The dictionary structure we have implemented with trie is **word (key) -> set of tweet_ids (value)**. The reason we have used a set data structure as value is because it allows us to perform union, intersection operations efficiently. Such operations help to find the result of query such as find all the tweets that mention trump and impeachment efficiently.

Example: **trie[prefix='impeach').intersection(trie['trump'])** gives tweet_ids where the word trump and prefix impeach is used. Similar operations can be used to find different combinations of words.

Initially construction of trie requires us to go through all the rows, tokenize the content column and insert word as key and create/update set with tweet_id. Without trie, if we are to iterate over the whole dataset to find occurence, it takes abou**t 5 minutes** to search the word whereas with trie it only takes a few milliseconds.

Google's trie implementation called pygtrie was used.

b. **Word2Vec and Doc2Vec**

Gensim was used to train word2vec and doc2vec model in CBOW mode. The reason CBOW was used is because CBOW takes less time compared to skip-gram though skip-gram is supposed to be more accurate with reasonable parameters. Whole dataset was used to train the model which resulted in vocabulary size of 562M words. Also, the size of serialized bin file of word2vec was 83MB and that of doc2vec was more than 1GB. Word2vec helped us to find similarity between words and detect biases present in the dataset, whereas doc2vec helped us to find similarity between tweets. Models were trained on i5 9th Gen 8 core CPU all the cores were deployed during the training process of this model. During the training process CPU throttled to its maximum capacity of 4.2Ghz.

.

We also used google's pre-trained word2vec model.

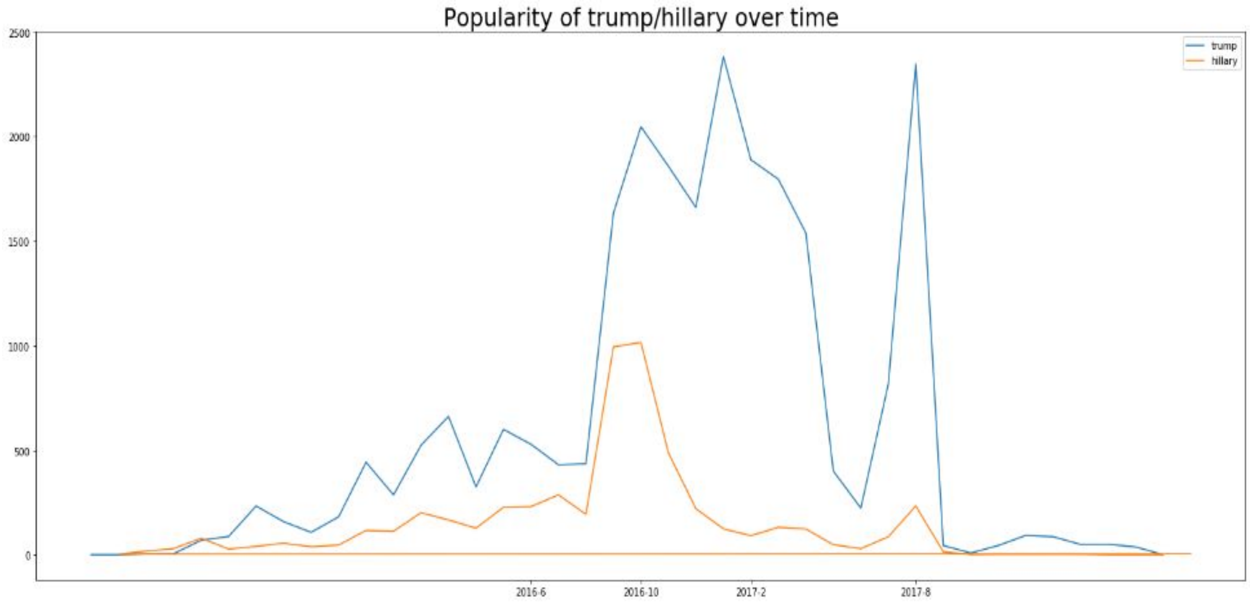
Pros: The vector operations like king - man + woman = queen were more precisely approximated.
Cons: Even though some general gender and racial biases were detected, political biases were difficult to detect. The dataset given to us contains political biases, which is why training a word2vec model would be more helpful to detect biases specific to the dataset.

Output for king - man + woman on our word2vec model.

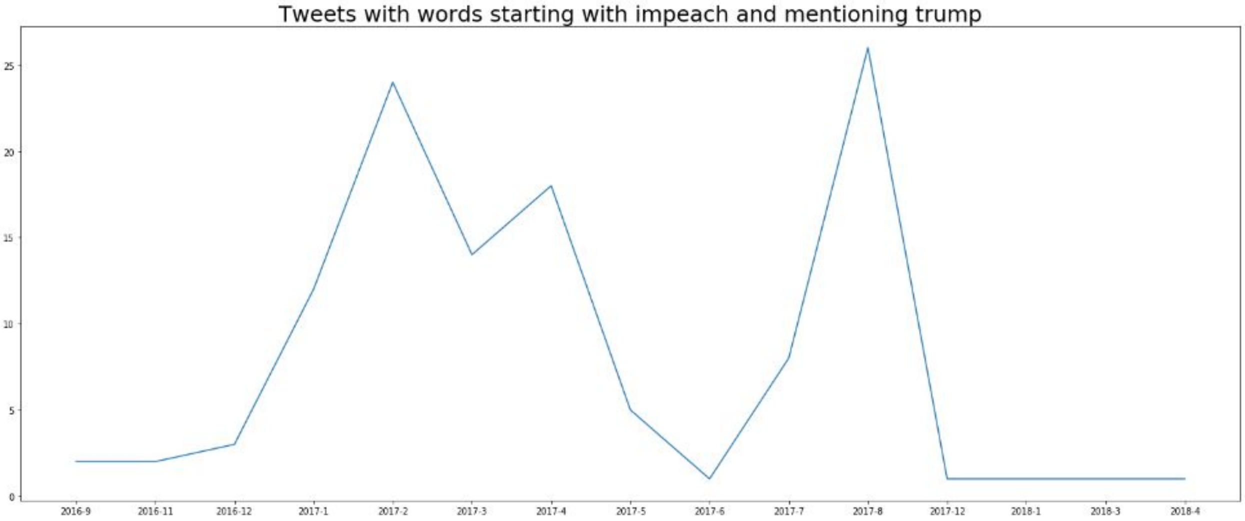
```
[('queen', 0.5206061005592346),
 ('david', 0.4740060269832611),
 ('legend', 0.463544636964798),
 ('president', 0.45934802293777466),
 ('president_obama', 0.44900429248809814),
 ('king_jr', 0.44154903292655945),
 ('champion', 0.4356076121330261),
 ('leader', 0.4292466342449188)]
```

Insight Visualizations

1. **Popularity of trump/hillary over time**



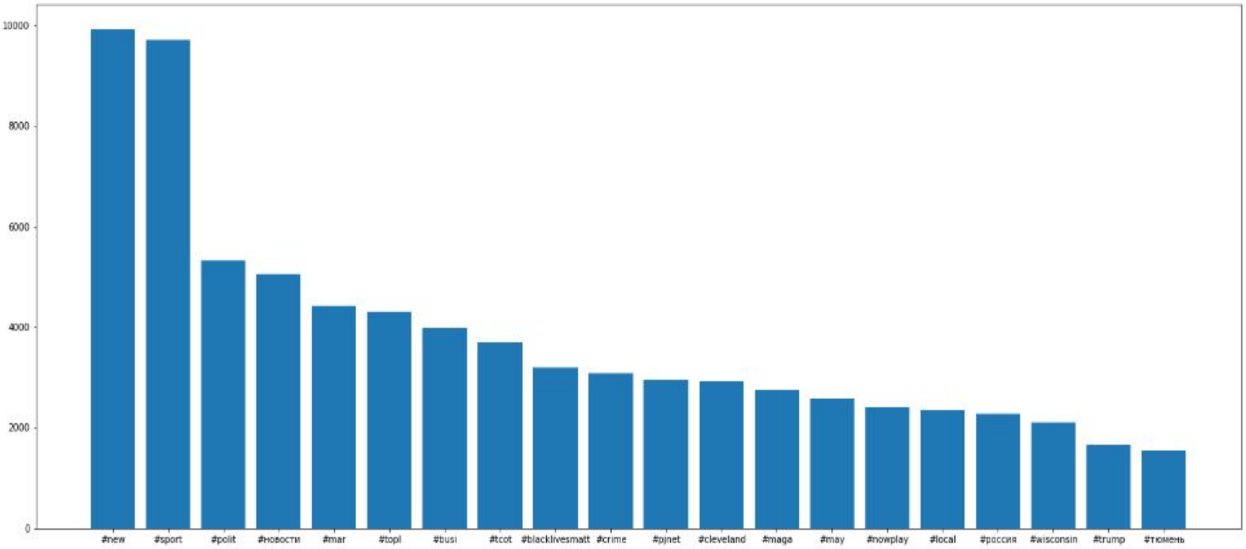
Before election date, the mention of trump is more than hillary because of the fact that the data represents russian meddlers who were supposed to support trump (and also trump is quite provocative) . After elections, hillary’s mention on twitter drastically reduces which is quite obvious. But in the month of august 2017, the mention of trump increases drastically. In august 2017, the russian investigation reached its prime; grand jury was called for the russian meddling investigation, which maybe the reason people are talking about trump a lot.



Talks about impeachment and trump also increased in the month of august 2017. It can be said that august 2017 was quite difficult for trump to handle.

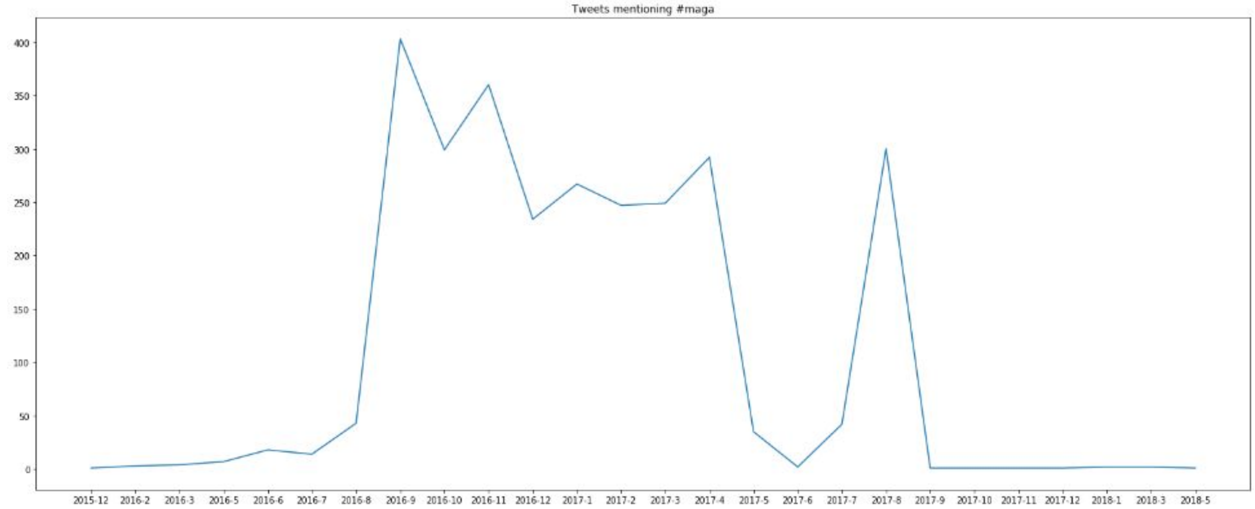
We used trie to generate such insight. Trie(prefix=‘impeach’).intersection(Trie[=‘trump’]) gives all the tweet_id where trump and impeach has been mentioned, and we make assumption that such tweets are discussing something about impeachment of trump.

2. **Popular hashtags**



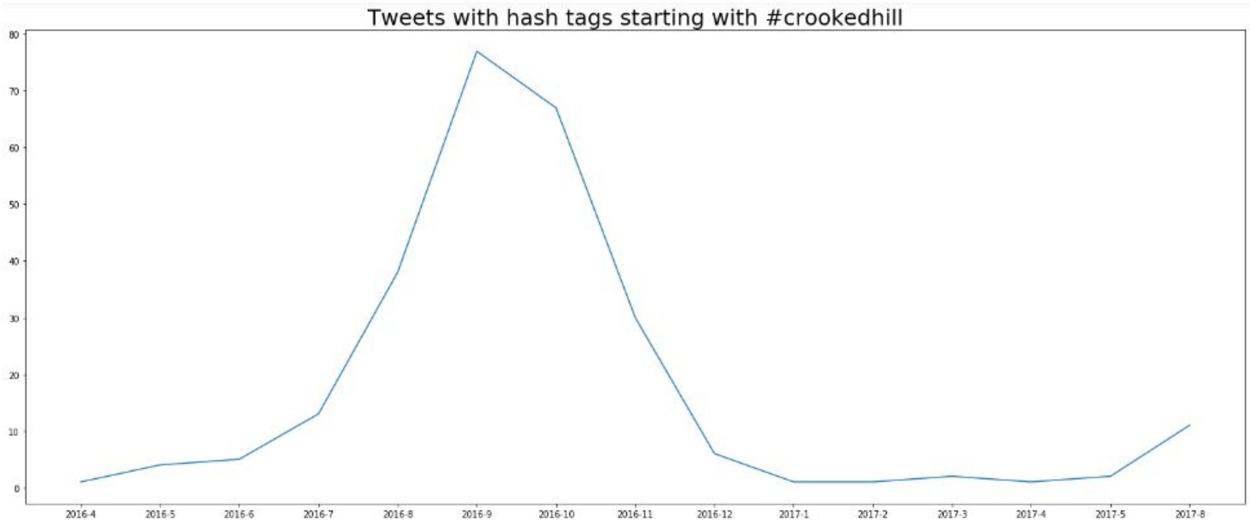
The most popular hashtag are news and sports. We think this contradicts the claim that the dataset represents russian meddlers because the popular topic seems to be sport and news rather controversial hashtags like #maga (Make America Great Again) and #tcot (Top Conservatives On Twitter). Or, maybe they talked about general topic to camouflage? Several hypotheses can be made but difficult to prove anything concrete.

3. **Tweets mentioning #maga (Make America Great Again)**



#maga was the main slogan of trump’s campaign. It can be seen #maga was very popular before the election which we think was used extensively to influence twitter users. The mention of #maga increased again in august of 2017 and it was the time when russian meddling investigation reached its peak. So, in order to influence/divert users, the topic #maga was made popular to show solidarity to trump during investigation.

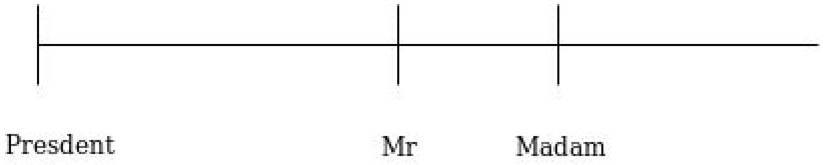
4. **Tweets mentioning hashtags which has prefix #crookedhill**



Variant for #crookedhillary was used which is why we queried for all keys starting with #crookedhill. It is quite obvious that #crookedhillary peaked just before election. It might be an attempt of trump supporters to ruin the image of hillary clinton.

5. **Biases**

   a. **Gender bias with word president**

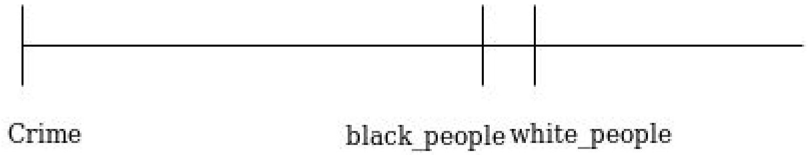


The above diagram shows the distance of Mr, Madam from the word president. We calculated the distance between president vector, mr vector and madam vector. **Distance(President, Mr) = 0.531**; **Distance(President, Madam) = 0.758**. This shows the word president is more related to male title and female title. Also, if we were to write an algorithm to generate sentences based on distances of these vectors, it would associate Mr with president with higher probability unless human intervention is applied.

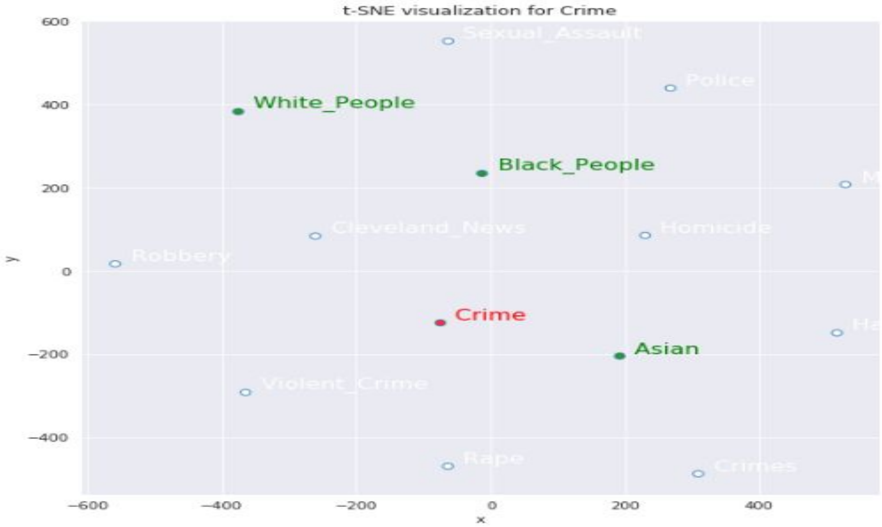
Some other example of gender bias are -

- **Cop;** distance('cop','man') = 0.59, distance('cop','woman') = 0.63. Cop is closed to man than woman
- **Nurse;** distance('nurse','he') = 0.60, distance('nurse','she') = 0.54. Profession nurse is more closely related to female than male.

   b. **Racial bias**



The above diagram shows the distance of black_people, white_peole from the word crime.**Distance(crime, black_people) = 0.74**; **Distance(crime, white_people) = 0.82**. This shows the word crime is more related to afro race than caucassians.

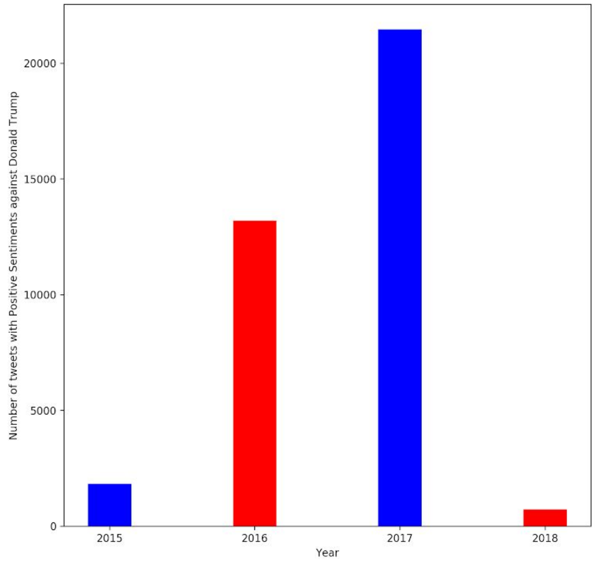


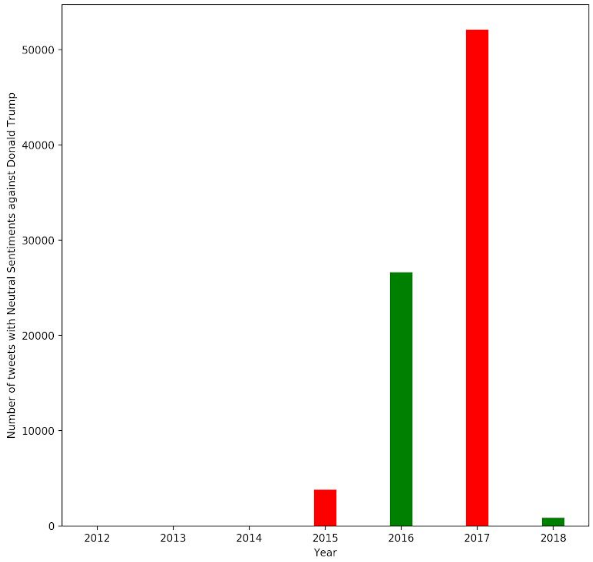
c. **Political bias**
   Distance(‘democrat’, ‘corrupt’) = 0.53
   Distance(‘republican’ ,’corrupt’) = 0.64
   Since the dataset represents trump supporters (right wings), bashing of democrats will be more than republicans and hence, democrat is more closely related to the word corrupt than the word republican.

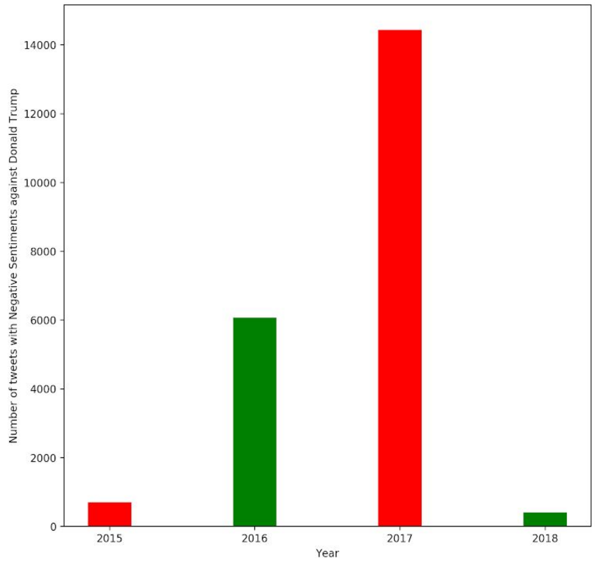
6. Sentiment Analysis

We filtered out tweets which were not about Donald Trump, Hillary Clinton and Barack Obama for sentiment analysis task. Doing so, removed tweets which were not required for sentiment analysis visualizations, giving us quite accurate visualizations.

1. Tweet Polarity when the it was about Donald Trump





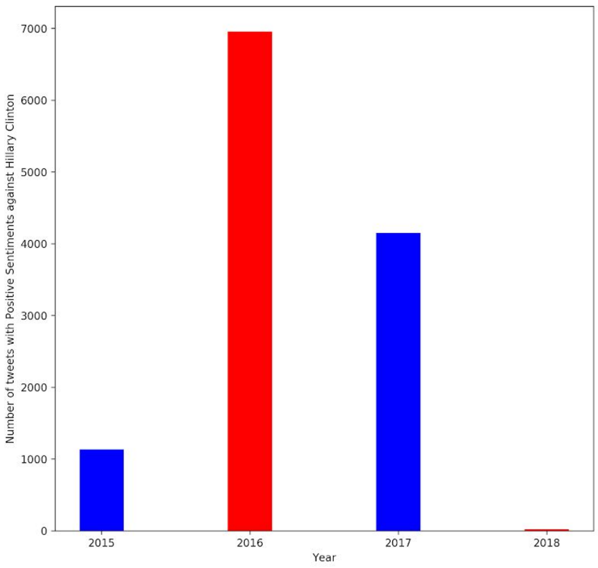


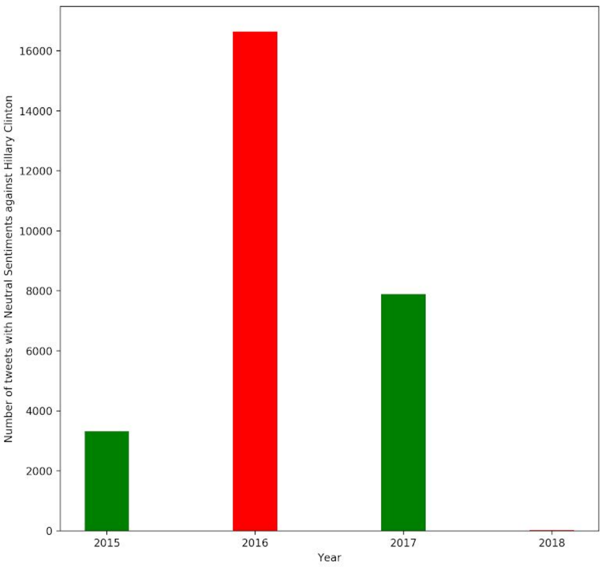
1. We have three cases here, positive neutral and negative polarity tweets. Polarity of a text is measured between -1 to 1 where -1 is negative meaning the text has a negative aspect to it(**words such hate, lie,bad, damage, fail etc**.), 0 meaning that the text is neutral (**words such as confident, selective, different, old etc**) and 1 meaning that

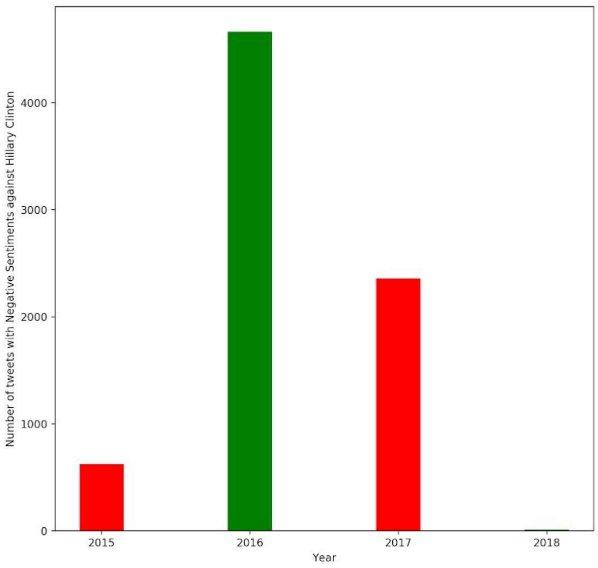
the text is positive in nature(**words such as amazing , beautiful, excellent, unique, courageous etc**).

2. Seeing the plots, tweets about Donald Trump increased significantly in 2017. Where number of neutral tweets were high, we have positive tweets and not too short the negative tweets. Ratio of Negative to Positive tweets for Donald Trump is **0.68 in the year 2017** and **0.46 in the year 2016**. Plots show that activity **tweet activity increased in 2017** as Donald Trump's presidency started on 20th Jan, 2017. After this, Trump has been involved in many controversies [1].This might be the reason for increase in number of tweets.
3. Talking about polarity Trump has been involved in many positive and negative things which is the reason for the polarity of the tweets[1][2]. Negative polarity suggests that Donald Trump's controversies were highlighted which were being defended by these twitter accounts posting about Trump, this can be inferred from tweets about Hillary Clinton.

2. Tweet polarity when it was about Hillary Clinton



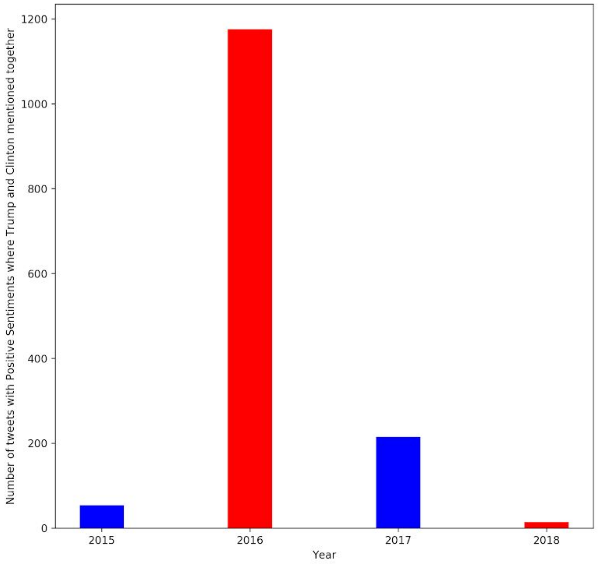


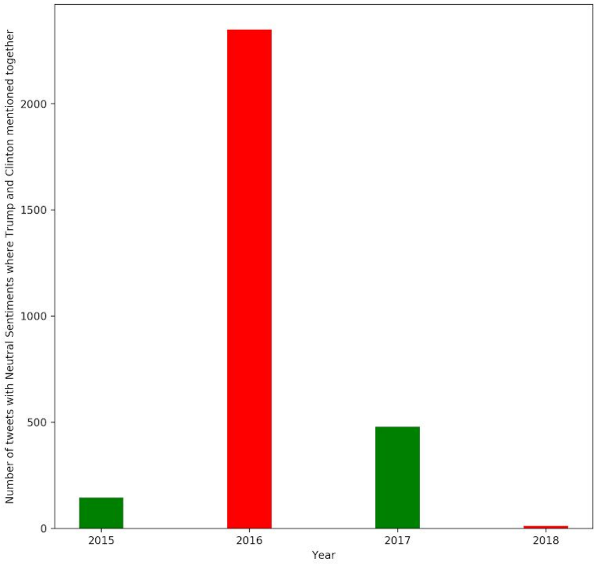


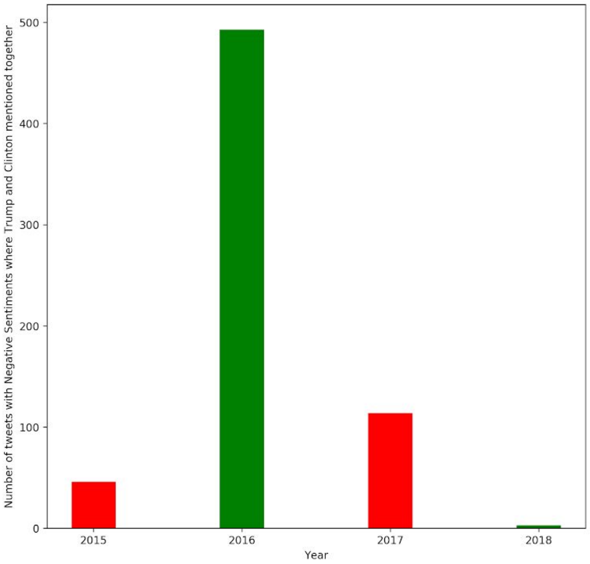
1. In the case of Hillary Clinton, tweet activity was highest in 2016 which is during election campaign. But still much less than Donald Trump which suggests active involvement of twitter in Donald Trump's presidency campaign. Also ratio of Negative to Positive tweets for Hillary Clinton is **0.70 in the year 2016** which is high when compared to Donald **Trump's 0.46 in 2016.** Which suggests that Hillary Clinton's controversies were being highlighted by these accounts which means these twitter handles were in favour of Trump.

2. Use of twitter and facebook was an integral part of Trump's presidency campaign[3]. People using twitter and facebook were shown **sponsored posts(paid business promotions)** on their newsfeed regarding positivity about trump and controversies about Hillary Clinton, hence the polarity. And according to the source [3], white middle age men voted for Trump more who were influenced by these sponsored posts.

3. Tweet polarity when it was about both Hillary Clinton and Donald Trump

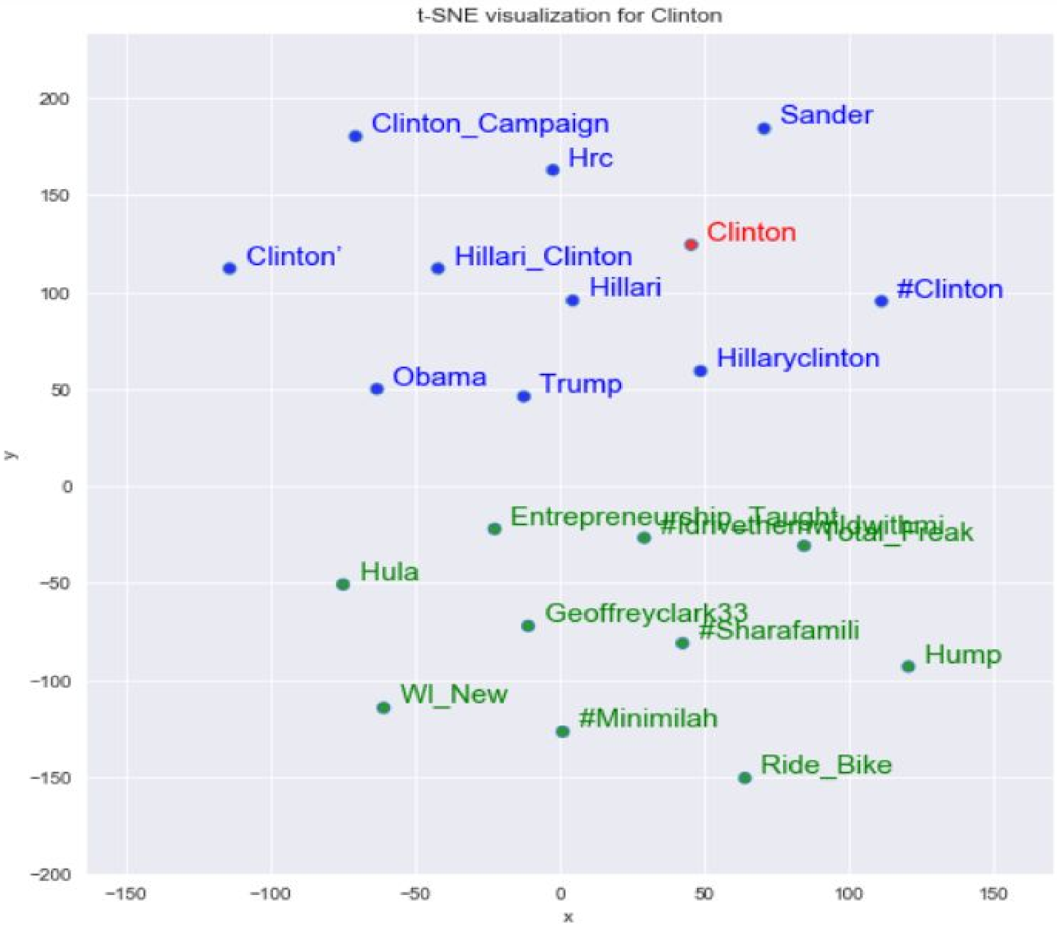




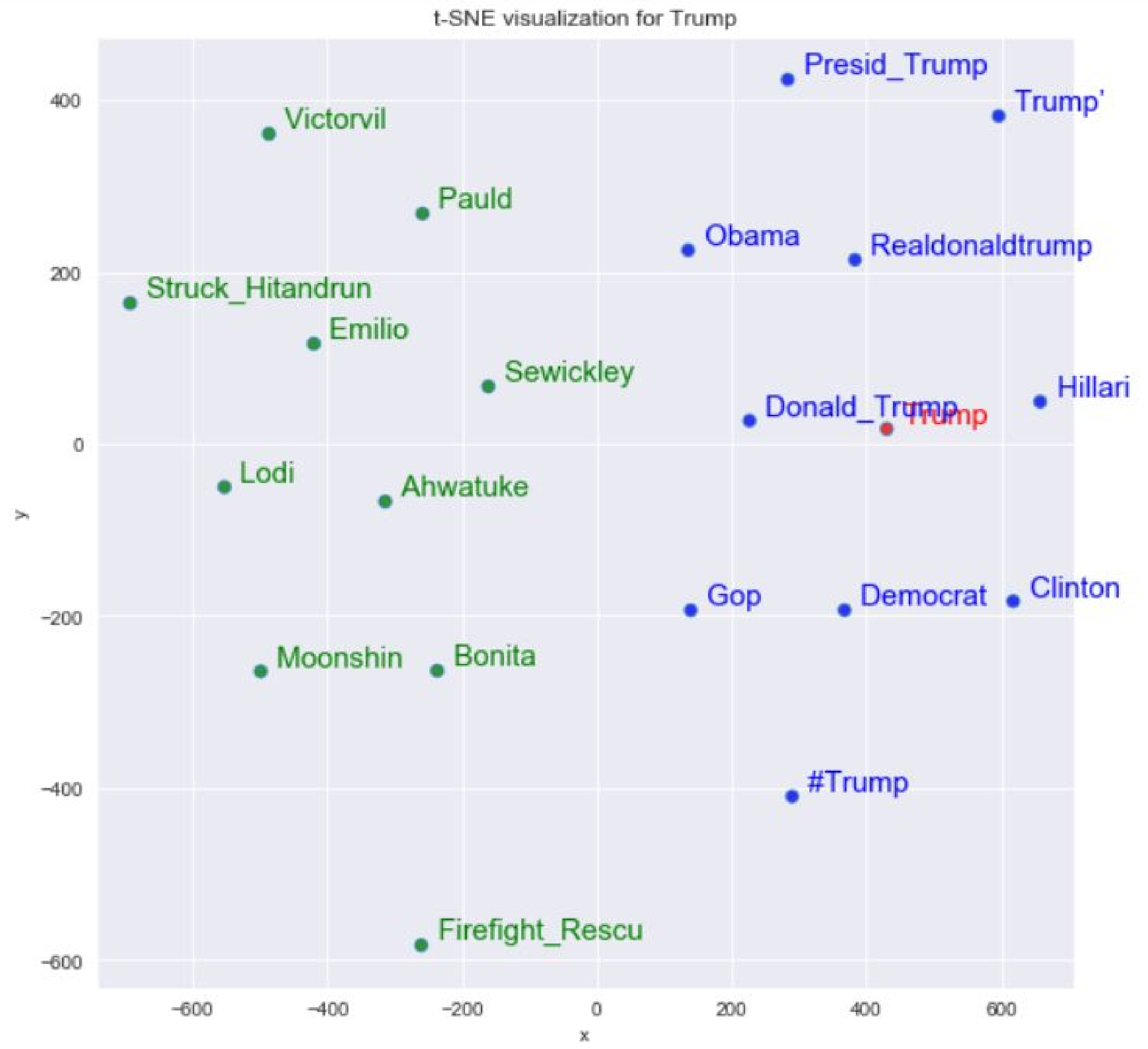


1. Donald Trump and Hillary Clinton were together on tweets from 2016 which is the time during the presidential campaign. It is quite possible that a comparison was done between the two which could be in favour of Trump as he won the elections. As mentioned in the source[3], social media played an important role in Trump’s presidential campaign.

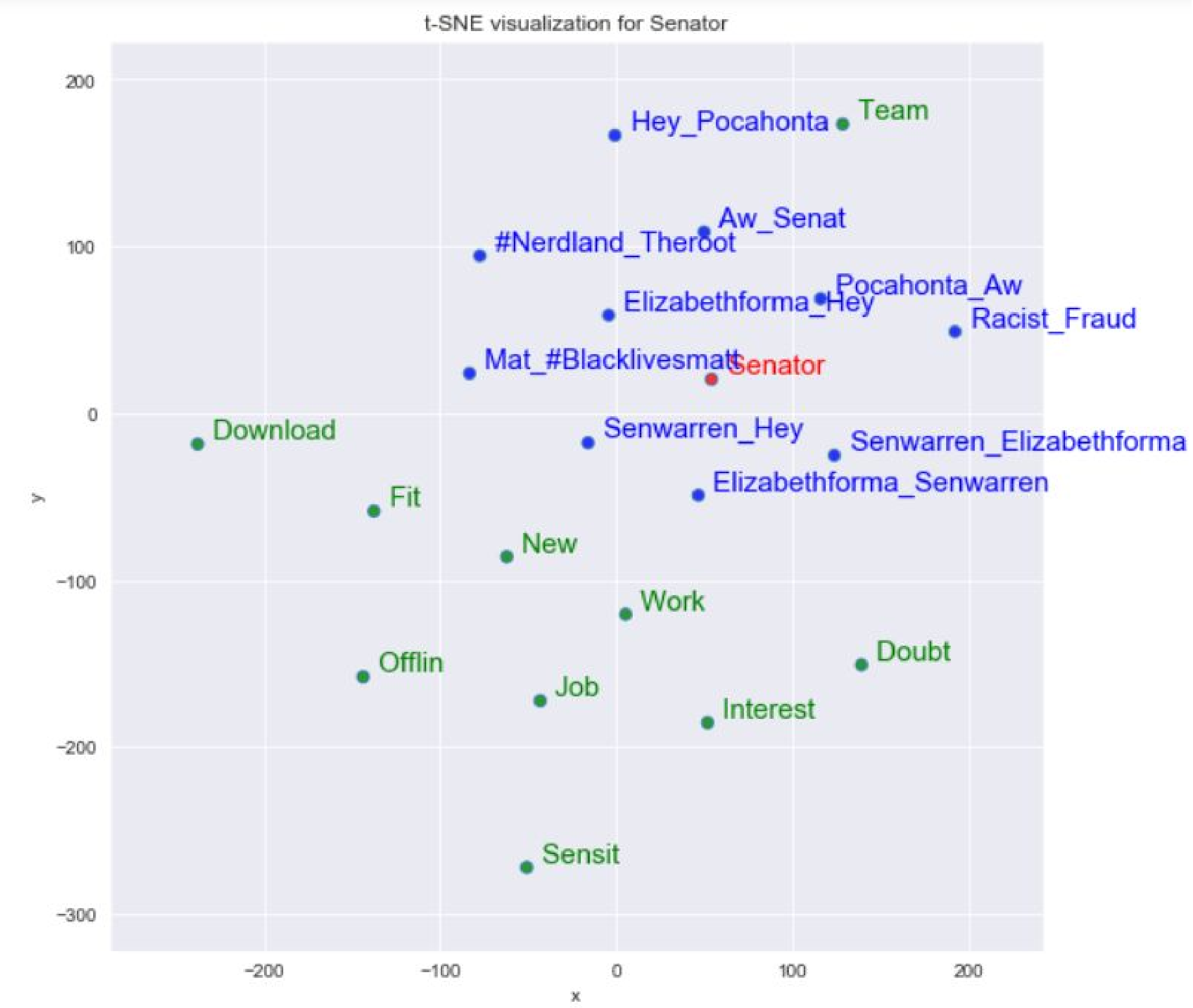
7. T-SNE Plots



1. This plot shows the 10 most **similar and dissimilar** words with respect to the word ‘Clinton’. Words in blue like Trump and Obama can be seen in the sense both of them being politician. This depicts the accuracy of our trained model. Words in green are the dissimilar words to the word Clinton.

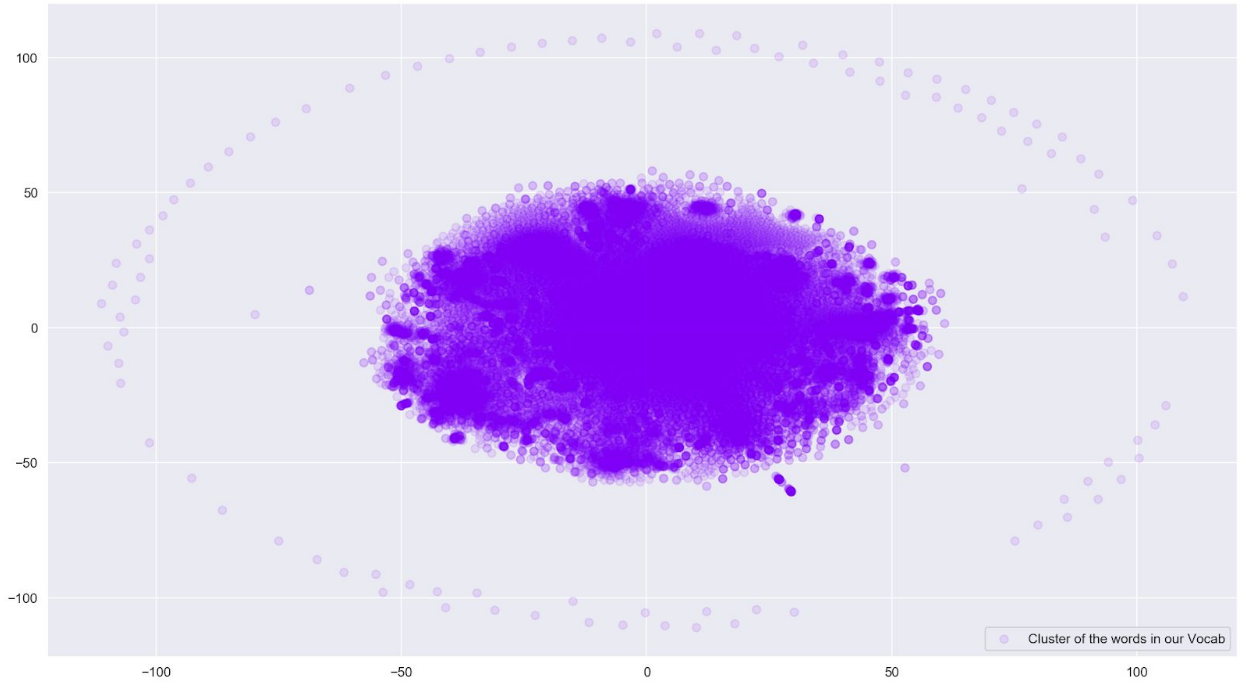


1. This plot shows the 10 most **similar and dissimilar** words with respect to the word 'Trump'. Words in blue like Obama and Clinton can be seen in the sense both of them being politician. This depicts the accuracy of our trained model. Words in green are dissimilar words.



1. This plot shows the 10 most **similar and dissimilar** words with respect to the word ‘Senator’. Words in blue like Elizabethforma and Pocahonta_Aw can be seen. **Senator Warren was mocked at by Trump who called her ‘Pocahontas’**[4] and that’s another controversy. This depicts the accuracy of our trained model. Words in green are dissimilar words such as job, work, team etc puts us in doubt regarding the working of Senators.

8. Word Cloud (Literal Cloud, Lol)



1. The plot shows all **562M words in our english Vocab as dots in this plot**. Density is very high in the centre due to the number of words and overlapping of the word clusters. But we have certain outliers surrounding the central dense cloud, which we think are badly mis-spelled words or non-english words.

Adherence to Rubric

1. We have successfully generated interesting insights depicting stories around US election and biases. Popularity of trump and clinton over time has been explained with hypothesis why the popularity peaked and why it decreased. We have also shown biases specific to the dataset like the biases against democrats and other biases like gender, racial biases. Also, the trend of tweet sentiments relating to trump and clinton has been explained.

2. As part of data modelling, we have used **trie** and **word2vec** for different purposes. We used trie to find the occurence of words and also reduce the memory usage compared to hashtables. We structured trie in such a way that allowed us to query the list of tweets that has occurence of multiple words together. We successfully trained word2vec model on the whole dataset and used it for bias detection and clustering.

3. As per the code shared, all the visualization should be exactly reproduced. The dataset is divided to parallelize to cleaning tasks such as lemmatization, stop words and url removal; so for better result make sure cleaning is run.
4. Looking at the paper "**Mitigating Gender Bias in Natural Language Processing: Literature Review** " by **Tony Sun , Andrew Gaut , Shirlyn Tang , Yuxin Huang, Mai ElSherief, Jieyu Zhao , Diba Mirza , Elizabeth Belding , Kai-Wei Chang, and William Yang Wang**

    1. Our fifth visualization is about Bias in our dataset where we found Gender Bias where Mr was near in distance to president than Mrs in other words our model assumed this analogy which can be seen in Table 1 on Page 2 of the paper. This concept is also called **Representation Bias.**
    2. Not just Mr and Mrs "President" also, He and Nurse, She and Nurse where model just assumed the analogy and distance between He and Nurse was high compared to She and Nurse.
    3. We think that this problem can be overcome by **Debiasing Gender in Word Embeddings.** This should work because our model is based on vector and their distances calculated in Euclidean Space. Using this method distance between He and She will reduce the distance between them and making them similar.

References

[1] https://thehill.com/homenews/campaign/366336-the-memo-the-top-10-trump-controversies-of-2017

[2] https://www.washingtonpost.com/opinions/the-10-best-things-trump-has-done-in-his-first-year-in-office/2017/12/27/c79ce93c-ea7e-11e7-9f92-10a2203f6c8d_story.html

[3] https://medium.com/rta902/year-one-how-donald-trump-used-social-media-to-win-and-maintain-the-presidency-fef7f7175d2c

[4] https://www.reuters.com/article/us-usa-politics-warren/senator-warren-mocked-by-trump-as-pocahontas-says-dna-test-backs-her-ancestry-idUSKCN1MP1I0

[5] Our Trained Models: https://drive.google.com/open?id=1qwZTj0tChe7X1WQyiVAhWxI42Oyd048w