CSC - 501

REPORT (ASSIGNMENT – 4)

TEXT DATA

SUBMITTED BY - GROUP B

Kapoor, Nikita	V00949256
Malik, Mona	V00935224
Vemula, Sannath Reddy	V00949217

Submitted To - Prof. Sean Chester

schester@uvic.ca

SECTION - 1

Data Modelling

We are given the raw text data on nearly 3 million tweets sent from Twitter handles over 6 years (February 2012 - May 2018) in the form of 13 csv files. Each csv file contains 21 columns which gives the important details of each tweet like the language, published date, twitter handle of author.

Data Cleaning:

We considered only the tweets written in English, Russian and Italian as these are the top 3 languages used in the tweets and they also cover 95% of the given data. Cleaned the raw data by filtering the rows with condition on column 'Language' (as language='English', 'Russian' or 'Italian') and further cleaning is done by text pre-processing.

Text pre-processing:

The given data needs to be filtered and pre-processed as it contains noisy data so, we considered removing hyperlinks, special characters, punctuations, stop-words, emojis, and numbers. The tweets (after cleaning the noisy data), emojis with description and hashtags are stored in a text file so that they can be accessed easily while working on insights.

Word Embeddings:

Word2Vector is the numerical representation of the words, which can be interpreted by the machine for text analysis.

We have implemented word2vector to create a trained model for the full corpus (cleaned dataset) using genism library. With the corpus specific model generated, we have done sentiment analysis (stored as text file) that gives polarity and subjectivity for every tweet. Textblob is the library which was used to produce sentiment analysis.

Bigram and 1-skip bigram were implemented out of the corpus specific model.

TF-IDF:

Term Frequency Inverse Document Frequency is another way of representing words in numbers. It tells the importance of a word in corpus with respect to other words in context.

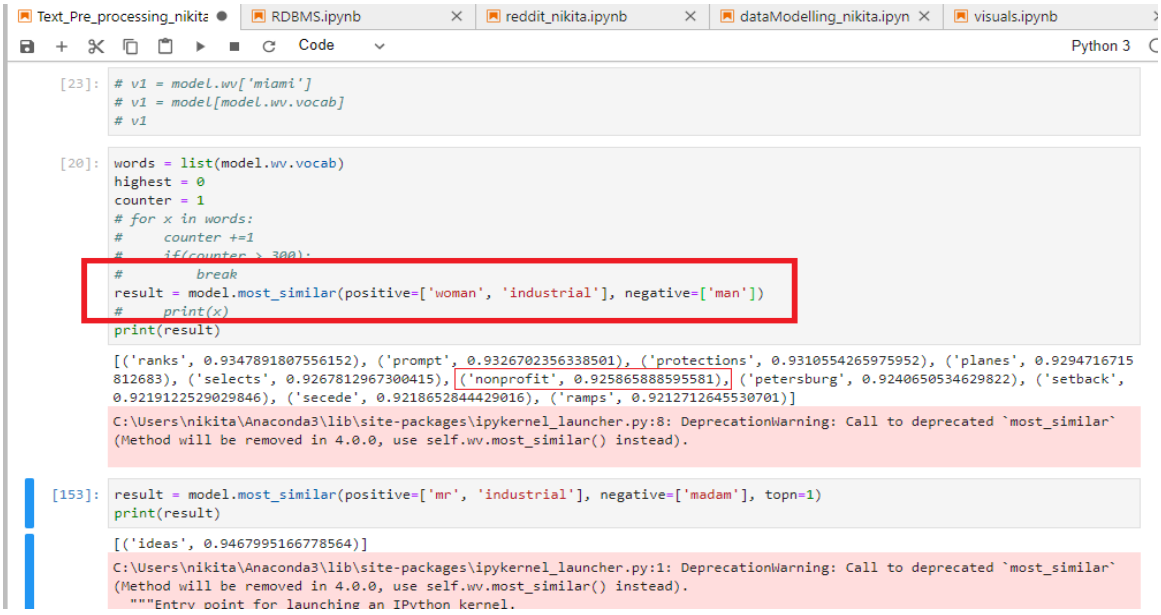
We have created TF-IDF vector for every tweet so that its possible to get the similarities between tweets. Below is the sample TF-IDF for a tweet

	TF-IDF
nedryun	0.359224
peep	0.330016
barely	0.317756
sitting	0.276971
trial	0.268678
youve	0.257540
mainstream	0.255078
heard	0.245645
corruption	0.240991
senator	0.230177
democrat	0.220584
media	0.174949
us	0.148469
from	0.148464
have	0.147187
we	0.144175
on	0.116593
and	0.108808
for	0.107407

Pre trained model:

Considered Google's word2vec model as pre trained model to check with biasing and other results. Below are some of the observations with pre-trained model,

Corpus-specific trained model:



Pre- trained model:

When compared to the trained model, pre-trained model fetches similar words with highest probability being 0.53(approx.). This doesn't specifically give the view whether the similar word(industrial) is close to (Industrial-man+woman).

Whereas, the trained model fetches the most similar words with probability of 0.9(approx.) which clearly tells that the similar words are much closer to the word which we are searching with.

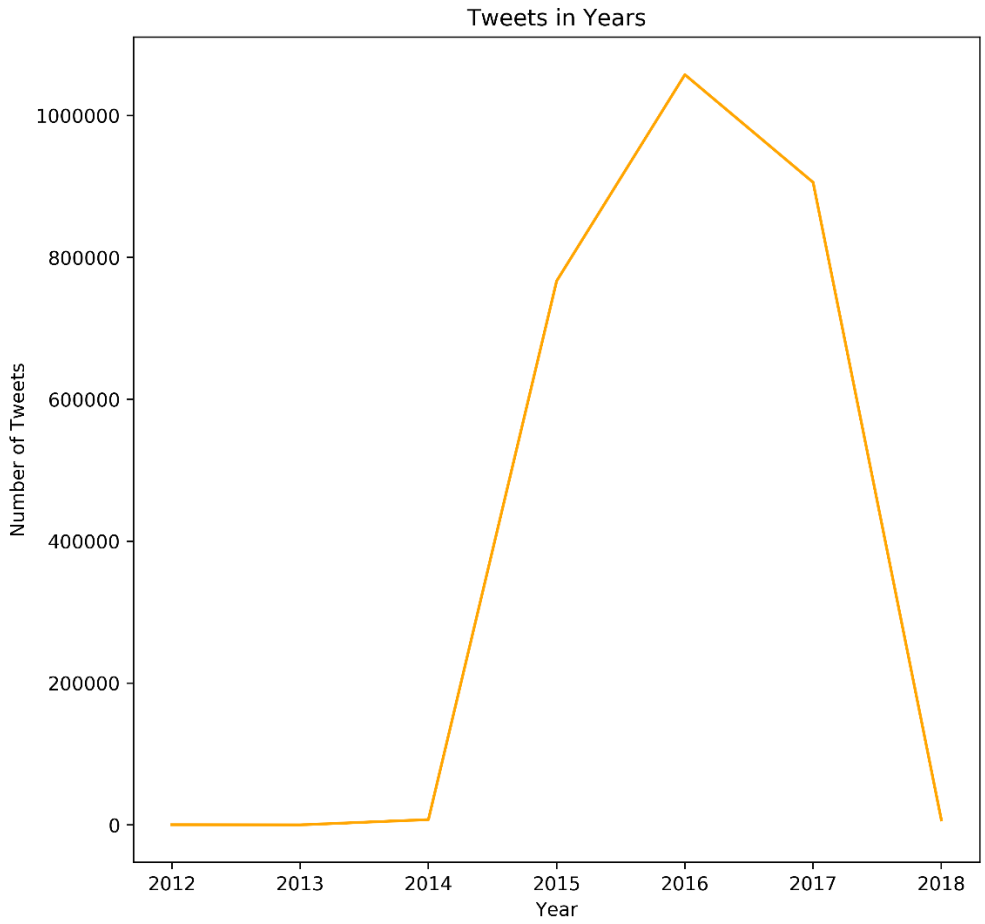
On the other hand, when looking at the words retrieved by pre-trained model, the words returned like industrial, shredded_warehouses appear to be similar in context even though the probabilities are less than 0.5 which is not the same case when we look at the results obtained by corpus trained model.

Biasing is clearly shown, as one of the words for Woman+industry-man is non-profit.

SECTION – 2

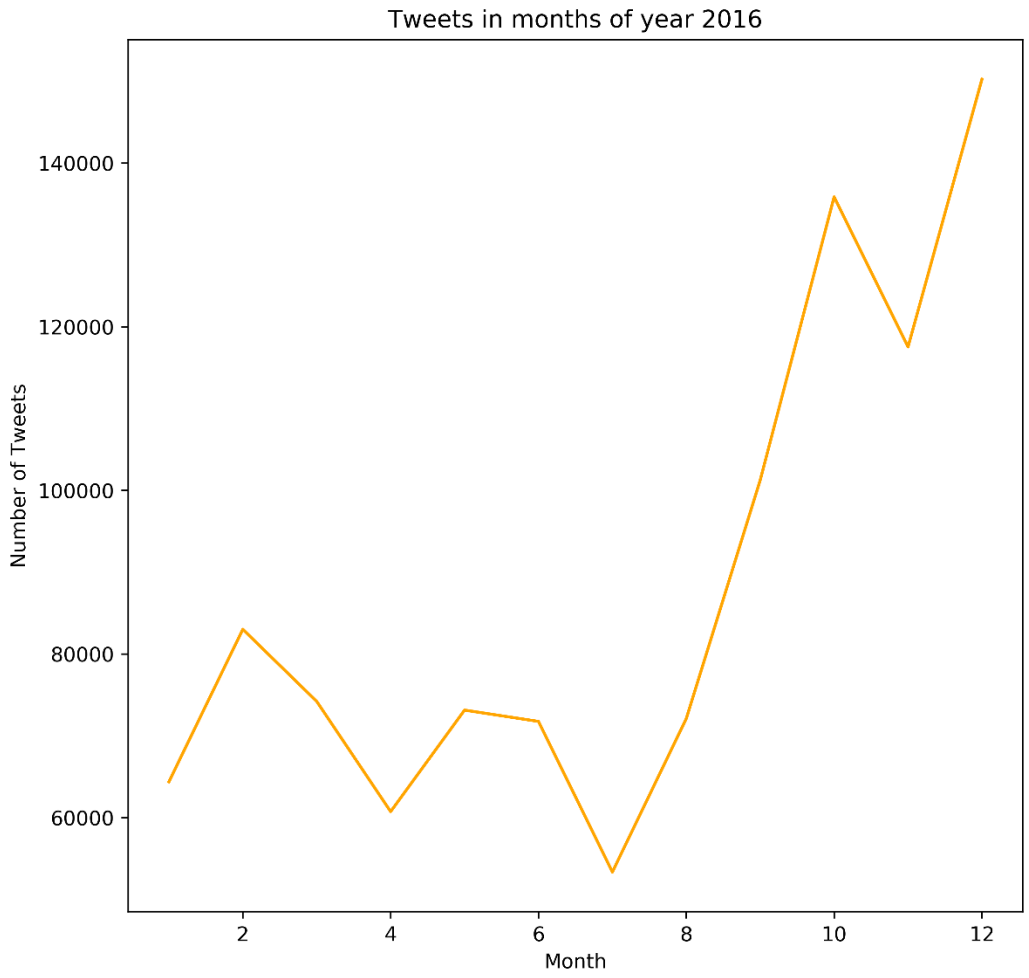
STORY TELLING

Visualizations:



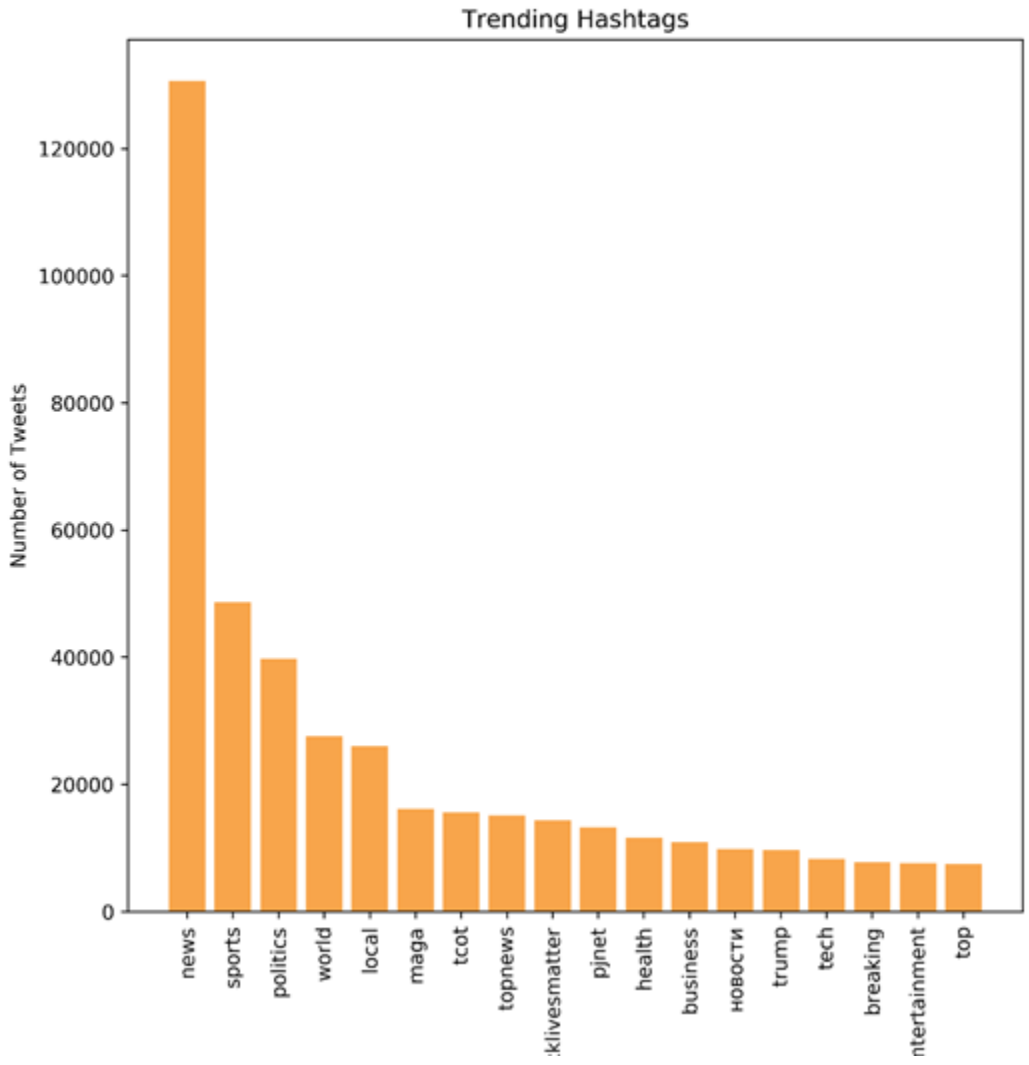
High usage of twitter in 2016 and the decline over the years

The graph depicts that there is a sudden increase in number of tweets after 2014, reaching the maximum in 2016. To see the popularity of twitter in 2016, we further plotted a graph for each month.



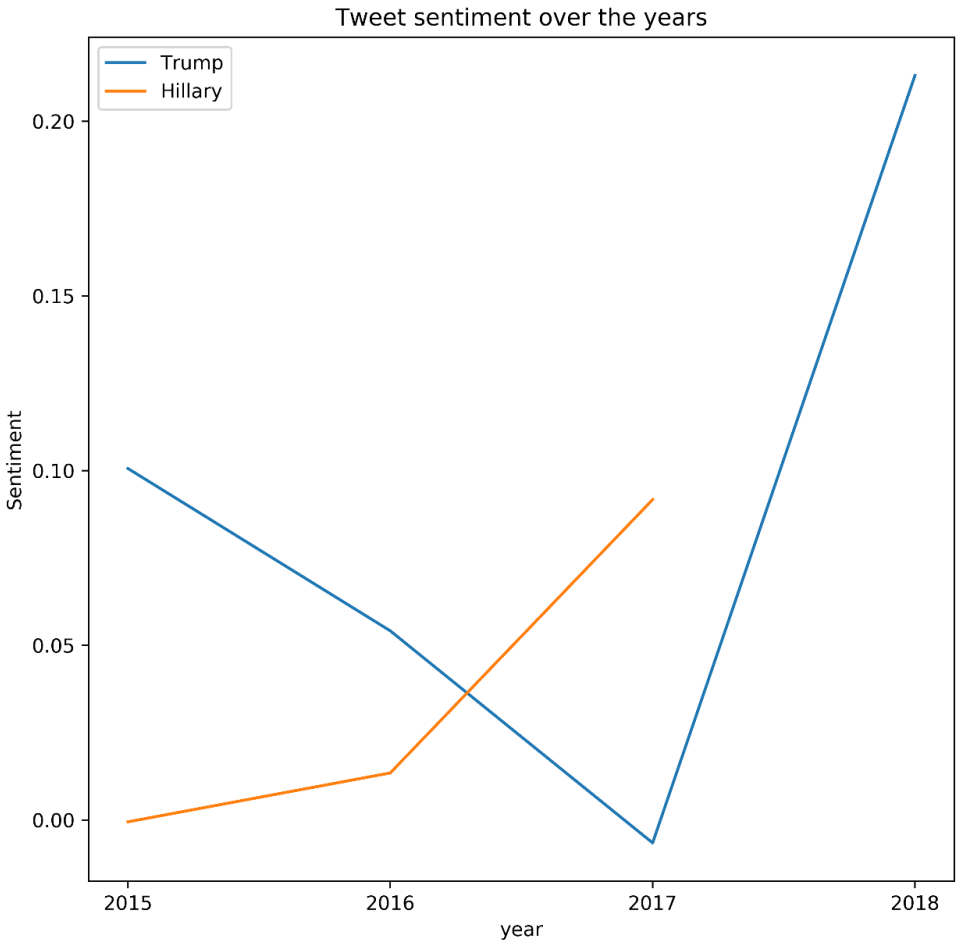
Sudden increase in number of tweets from July 2016

In **October (change month)** 2016 twitter was used extensively and maximum number of tweets were done in December 2016 which clearly states that US presidential election was the hot topic being discussed among the people worldwide. To validate this reason for the high usage of twitter, we checked popular hashtags used in tweets in 2016.



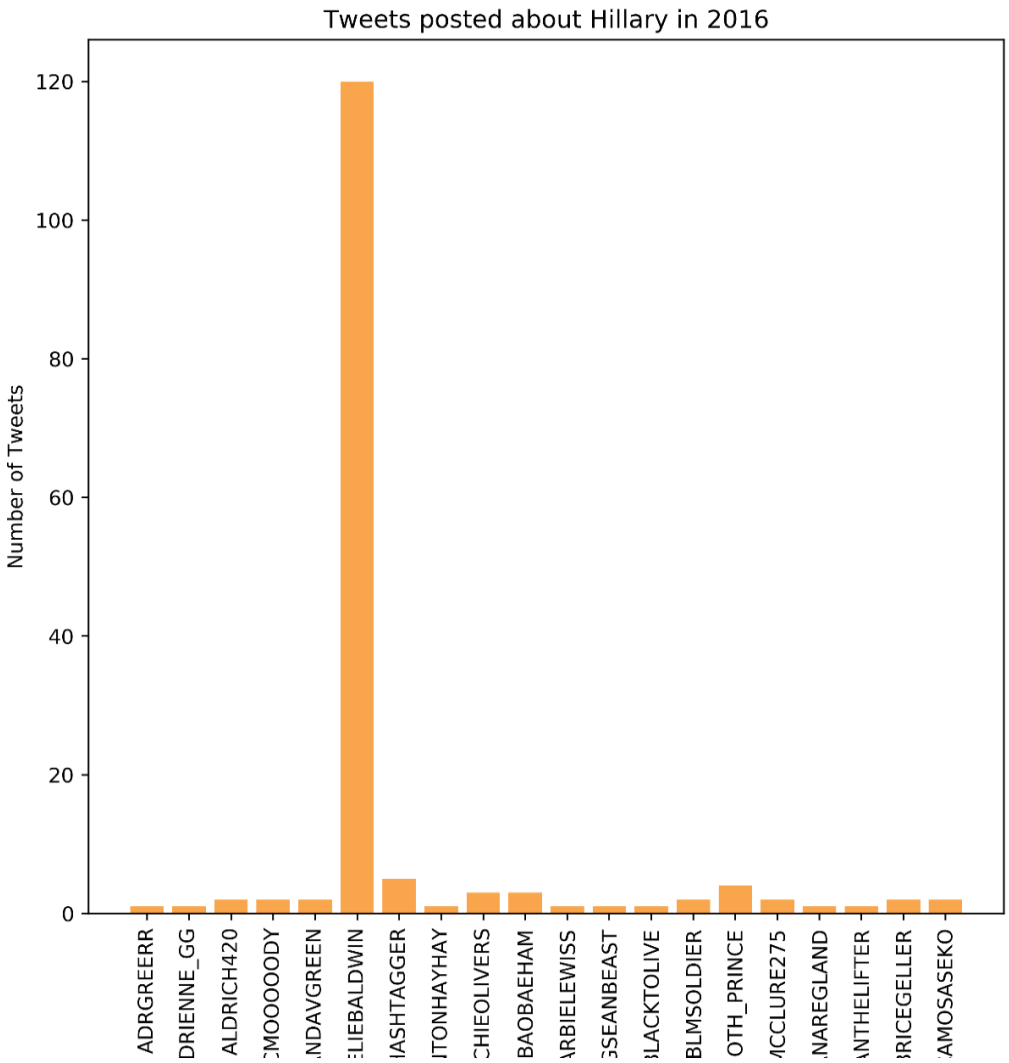
Presidential election's popularity reflected by the hashtags used in tweets in 2016

Hashtags are the words that gives the summary about the topic of tweet. We checked popular hashtags being used in year 2016 and the event related to it which was leading to sudden increase in usage of twitter.



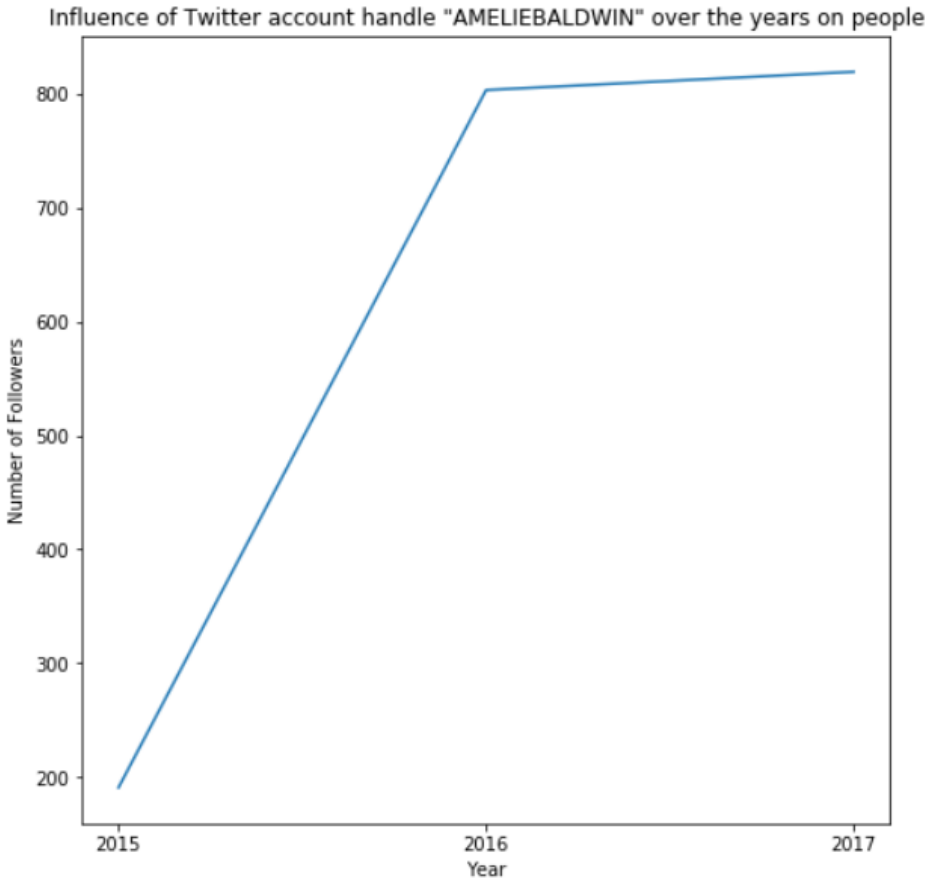
Positive sentiment for Trump and negative for Hillary in 2016

As the popularity of presidential elections increased, we checked the sentiment polarity of the tweets which were related to Hillary Clinton and Trump. This depicts that in 2016, people were positive about Trump and negative about Hillary which clearly explains Trump's victory in the election in 2016.



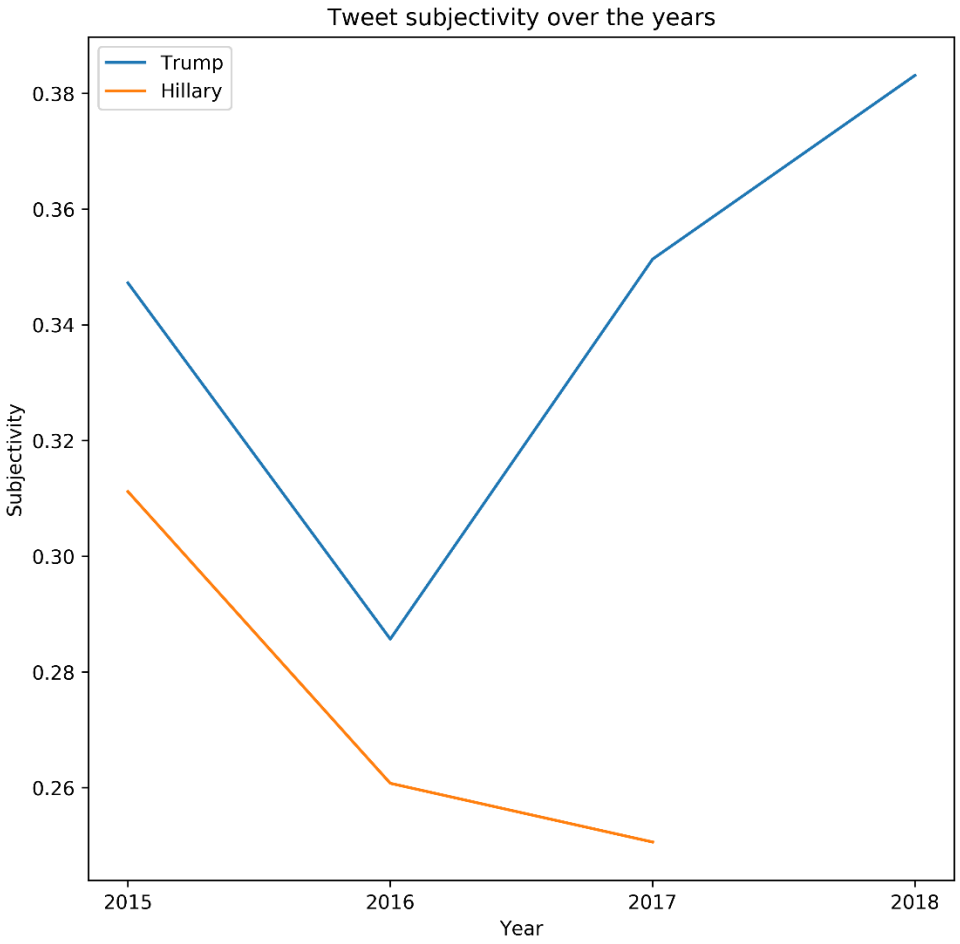
Tweet handle named "AMELIEBALDWIN" spreading the negativity for Hillary on twitter over the year 2016

There was a sudden increase in negativity for Hillary among people. We checked the number of twitter handles posting negative tweets over the year 2016.



Twitter handle named AMELIEBALDWIN's negative tweets increased influence on people

Despite of the creation of Twitter handle named "AMELIEBALDWIN" in **2015(change)**, it influenced many people over the years with its negative tweets and the account got popular at a quite fast rate as we see the increase in its followers.



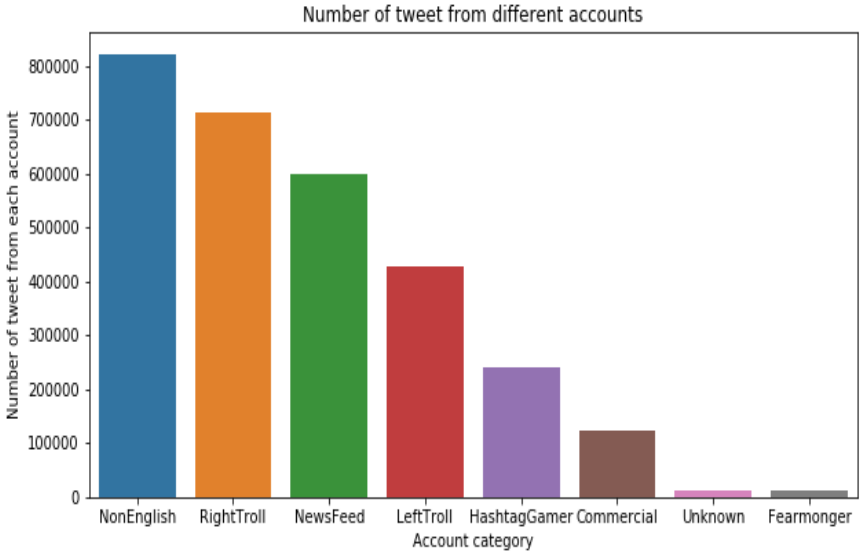
People have more opinions about Trump after 2016 Presidential election as compared to Hillary's stable image.

Subjective expressions are opinions that describe people's feelings towards a specific subject or topic whereas objective expressions are the facts.

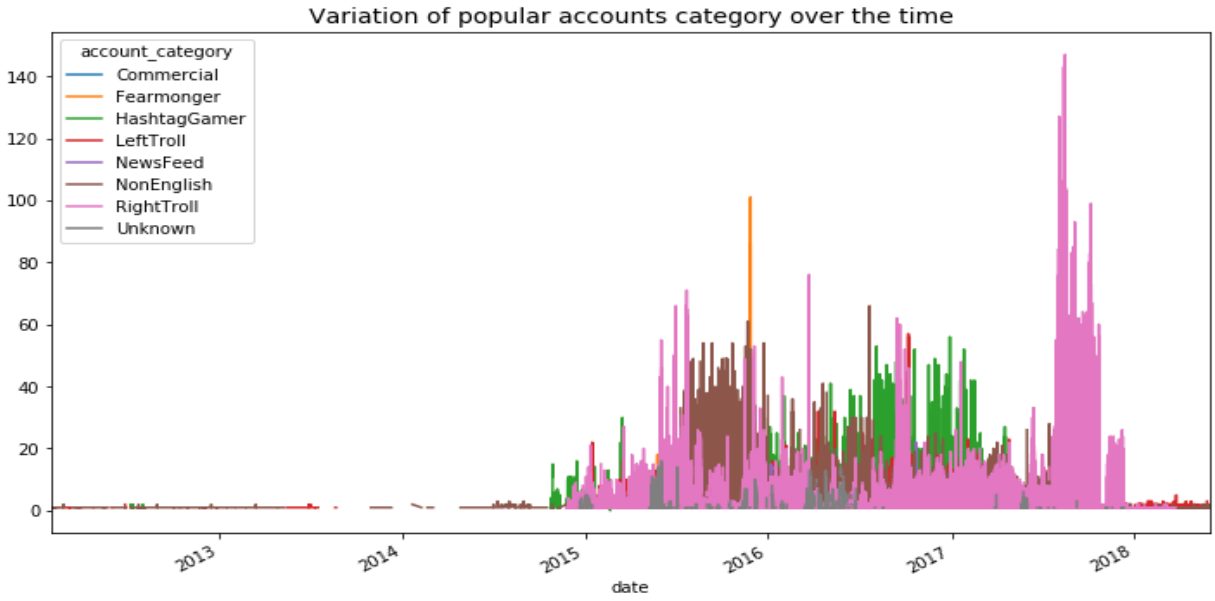
After the elections of 2016, it seems that many people have opinions for Trump which is not known whether it is positive or negative but can be proved in the upcoming elections whereas people have less opinions about Hillary which also depicts that people have same mindset for her as it was before.

Few other Insights:

INSIGHT #1: **Number of Tweet from different Troll Accounts and its variation over the time**




```
{'RightTroll': 711668,
 'NonEnglish': 820803,
 'Fearmonger': 11140,
 'LeftTroll': 427141,
 'Unknown': 13539,
 'HashtagGamer': 241786,
 'NewsFeed': 598226,
 'Commercial': 121904}
```

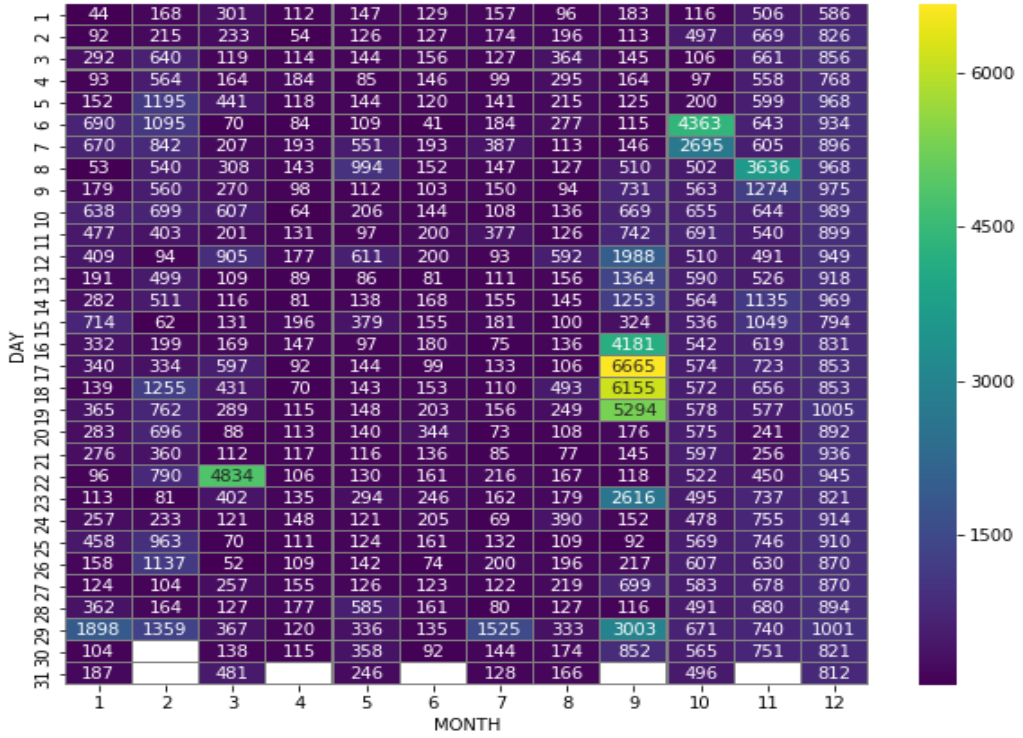


From the complete dataset we analysed that in the English tweets the most active account in tweets is Right Troll, followed by News Feed and Left Troll. From the second visualization, we analyzed that the Right Troll tweets has grown significantly from 2015 to 2018. Also, we can see News Feed was more active from 2016 – 2017 and commercial from mid 2015 – mid 2017

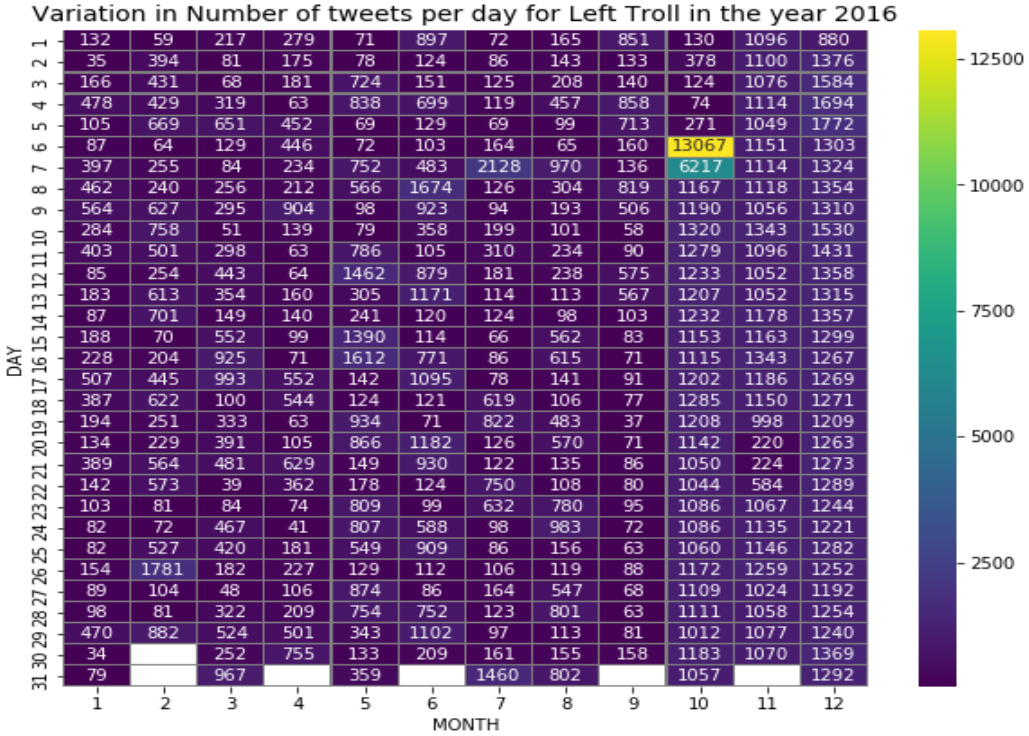
INSIGHT #2 : Left Troll Vs Right Troll(main Troll Account during Campaign)

2.1 Right & Left Troll accounts tweet count for the year 2016,2017,2018

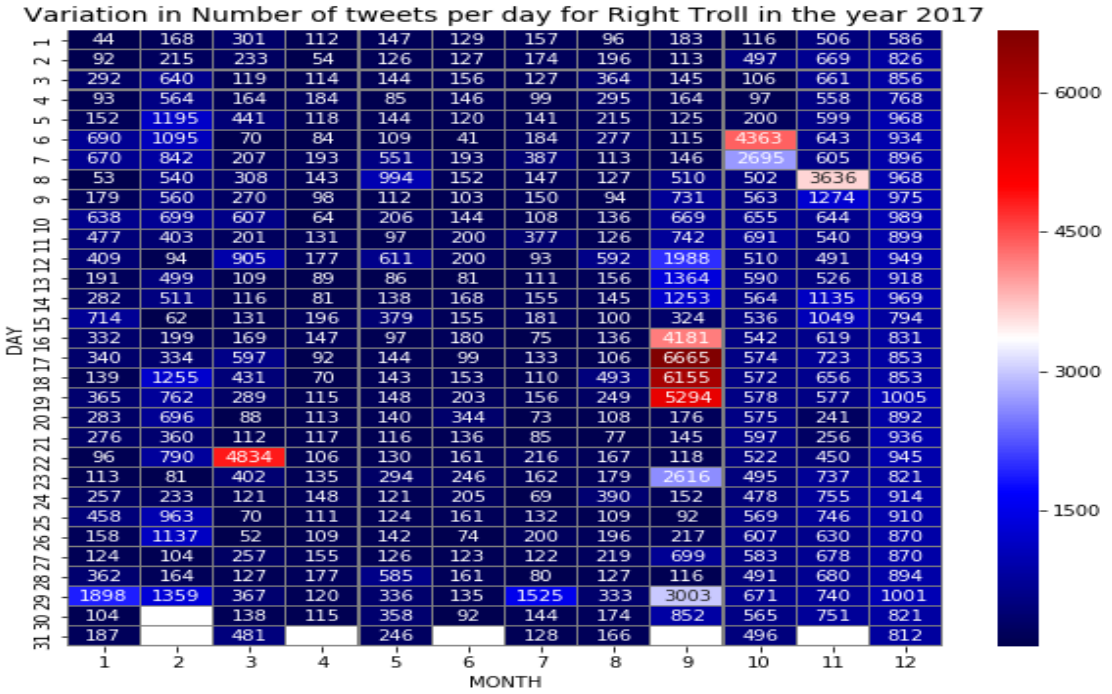
Variation in Number of tweets per day for Right Troll in the year 2016

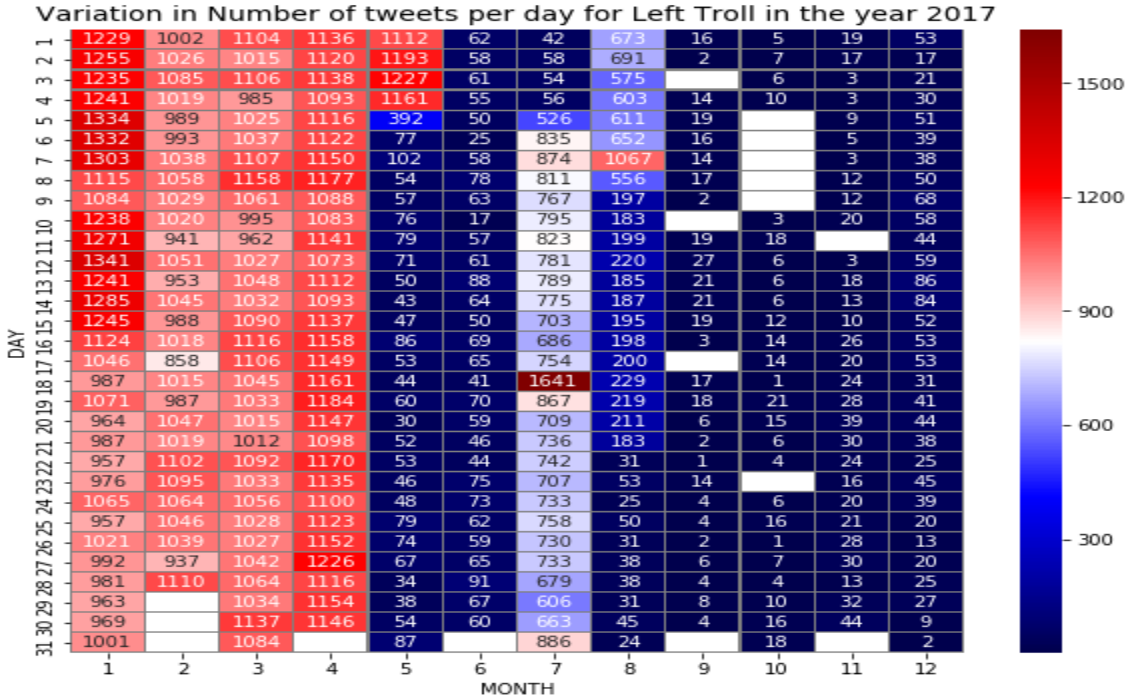


Variation in Number of tweets per day for Left Troll in the year 2016



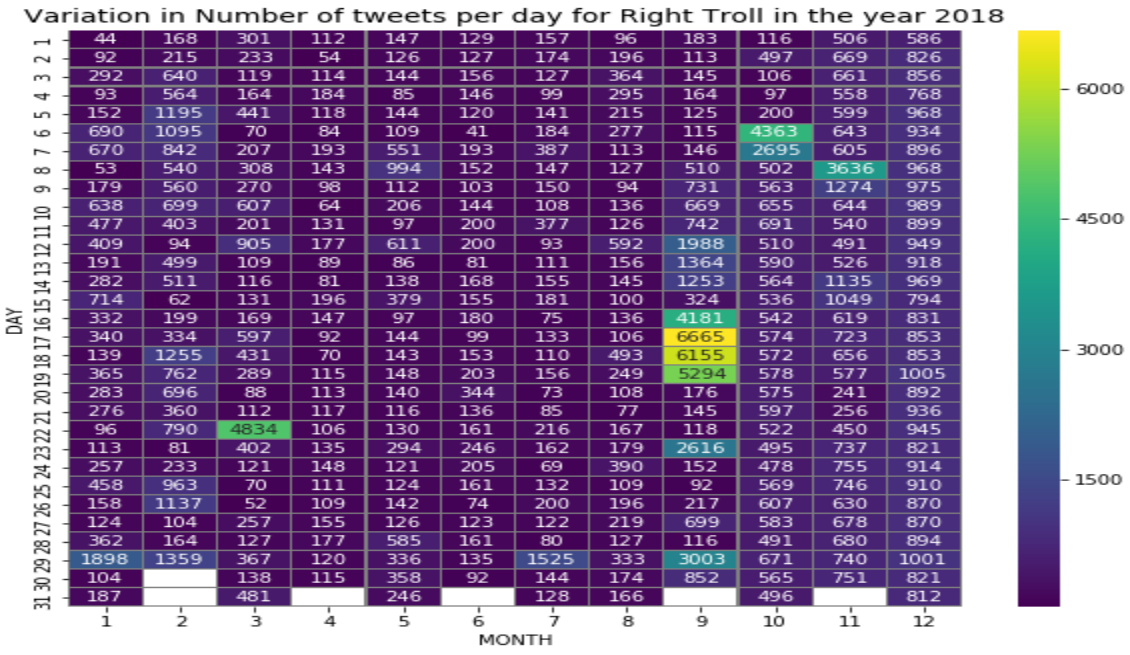
Left Troll and Left Troll were more active on Oct, Nov, Dec with maximum tweets on 6,7 Oct by Left Troll account. RightTroll was comparatively less active than Left Troll in the year 2016, was most active on 15-19 Sep. Its strange to note both left and right troll do not have any tweet on same days.

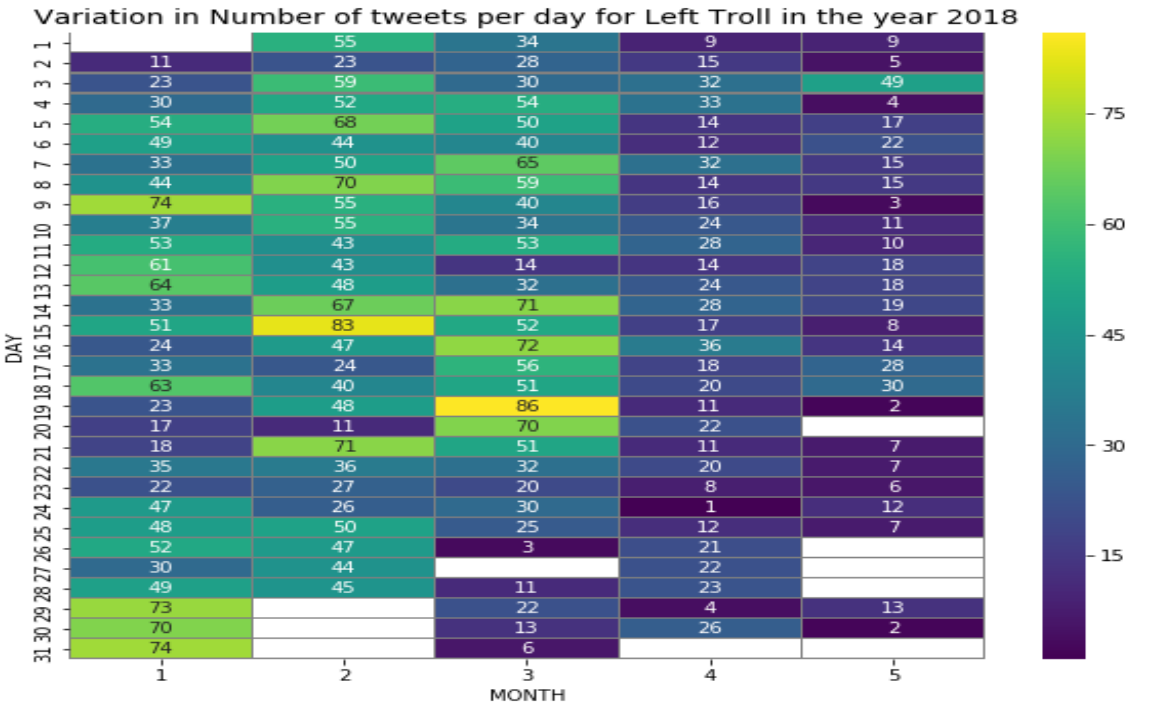




In the year 2017, Left Troll was a lot active in the initial 4 months and very less active in the last four months while Right Troll has hightest tweet in the end of year. It is strange to note that Right Troll have

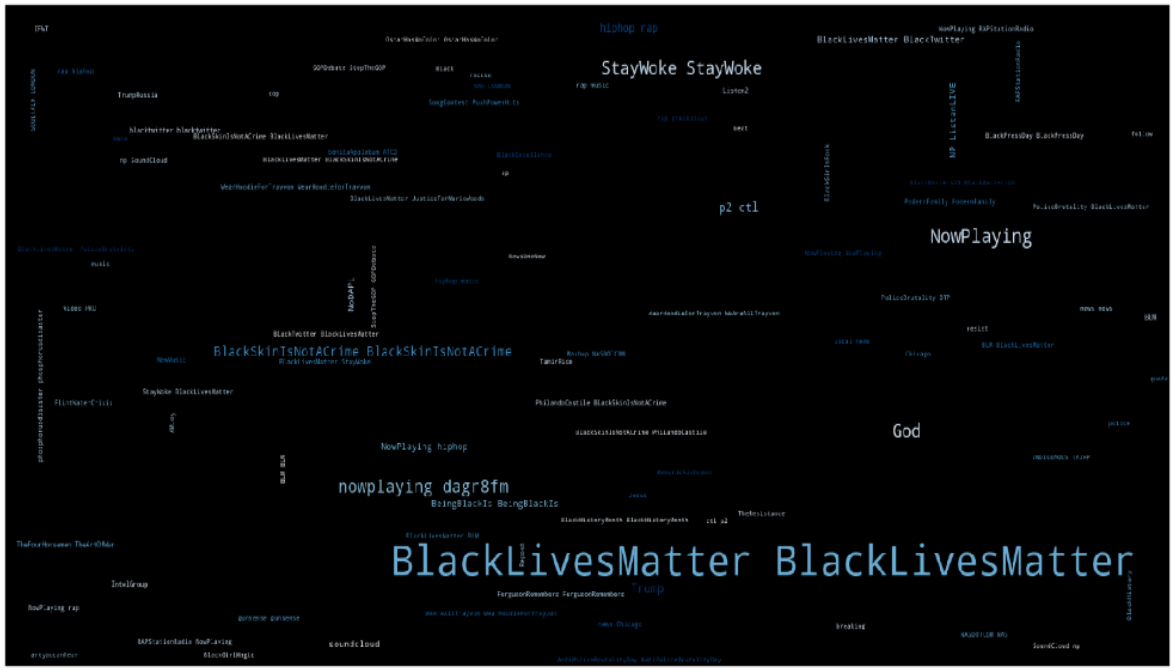
good number of tweets on 6 Nov 2017 while Left Troll have none. Right Troll is consistent with Tweets whereas Left Troll was only active in initial months and somewhat went active again on July too but rest not much active.





In the year 2018 left Troll was very inactive while Right Troll was significantly active, most active in the Last Four month of the year. Maximum Tweets are from 15-19 Sep by the Right Troll

2.2 Popular Hashtag's used by Left Troll Vs Right Troll



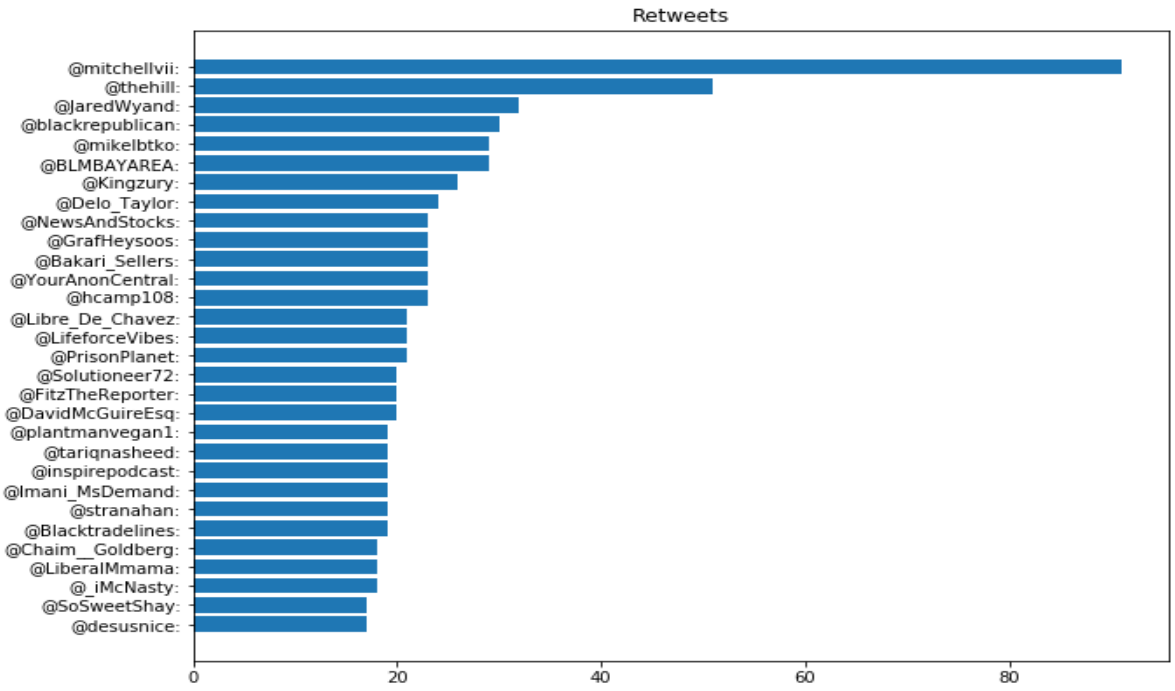
LEFT TROLL POPULAR HASHTAGS

We can see the popular Hashtags by Left Trolls are #BlackLivesMatter #Blackskinisnotacrime

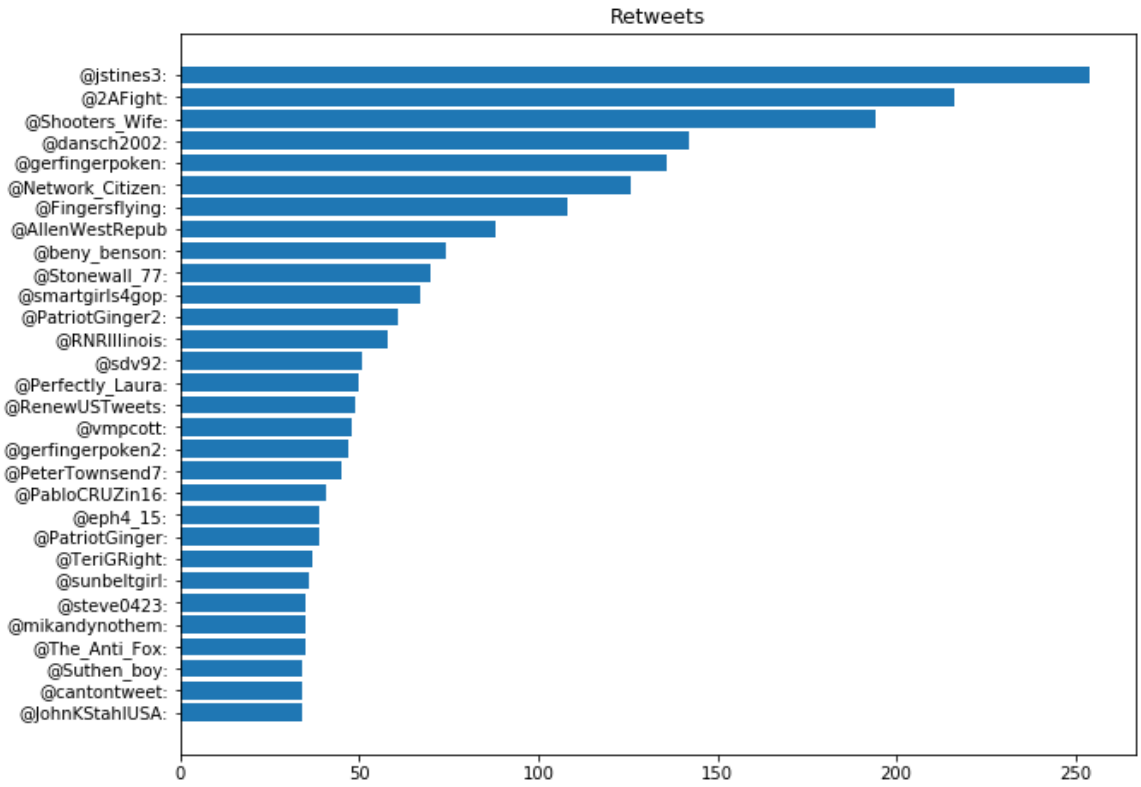
RIGHT TROLL POPULAR HASHTAGS

We can see the popular Hashtags for Left Trolls are #Top #FakeNews #Hillary #TrumTrain #MAGA #IslamKills

2.3 Popular Retweets used by Left Troll Vs Right Troll



Popular retweets for the Left Troll are @mitchellvithei,@Prison Planet,@ Blackrepublican , @thehill



Popular retweets for the right Troll are@jstine3,@2AFlight,@Network_citizen, @RenewUStweets and others

CONCLUSION

Right Troll and Left Troll are the meat of the agency's trolling campaign. Left Trolls often adopt the personae of Black Lives Matter activists trying to divide the Democratic Party and lower voter turnout while right Right Trolls is very active in American Politics

Topic Modelling:

LDA model (Latent Dirichlet allocation) is built with 5 different topics where each topic is a combination of keywords and each keyword contributes a certain weightage to the topic.

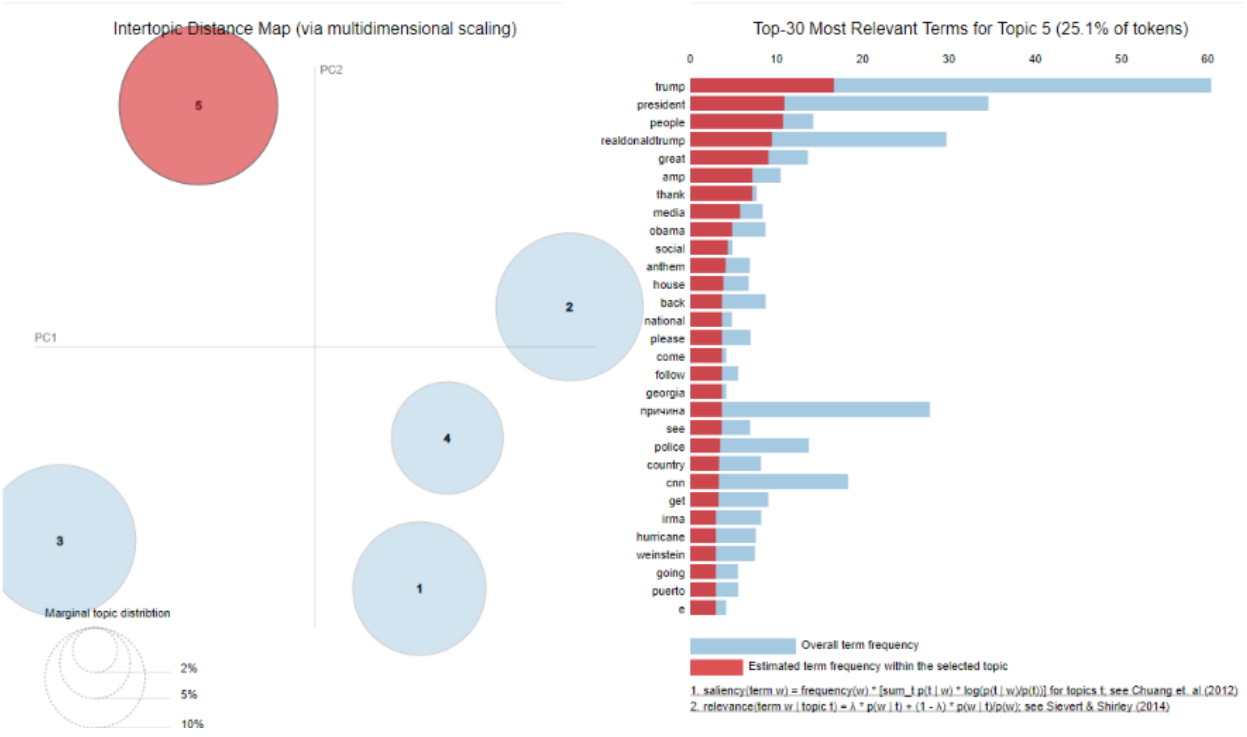
In the Visualization each bubble on the left-hand side plot represents a topic. The larger the bubble, the more prevalent is that topic. Right-hand shows the salient keywords that form the selected topic and their weightage.

Model perplexity and topic coherence provide a convenient measure to judge how good a given topic model is.

```
Perplexity:  -7.870451016183864
Coherence Score:  0.5757121463422921
```

Cons – This model takes a lot of time to build and visualize. The work is done only on 1000 tweets to get the comp utational measures.

Reference - https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/



Connection to research:

Mitigating Gender Bias in NLP:

This paper speaks about the biasing in NLP and gives an algorithm to mitigate the biasing. In this assignment we have corpus specific model and observed biasing with few words. When pre trained model of google or twitter is considered, it may lead to some other similar words. Below is the screenshot of biasing observed in the corpus specific trained model in comparison to google's pre trained model.

As said earlier in modelling, biasing is evident from the similar words returned.

```
# for x in words:
#     counter +=1
#     if(counter > 300):
#         break
result = model.most_similar(positive=['woman', 'industrial'], negative=['man'])
#     print(x)
print(result)
```

```
[('ranks', 0.9347891807556152), ('prompt', 0.9326702356338501), ('protections', 0.9310554265975952), ('planes', 0.9294716715812683), ('selects', 0.9267812967300415), ('nonprofit', 0.925865888595581), ('petersburg', 0.9240650534629822), ('setback', 0.9219122529029846), ('secede', 0.9218652844429016), ('ramps', 0.9212712645530701)]
```

```
C:\Users\nikita\Anaconda3\lib\site-packages\ipykernel_launcher.py:8: DeprecationWarning: Call to deprecated `most_similar` (Method will be removed in 4.0.0, use self.wv.most_similar() instead).
```

[153]:
```
result = model.most_similar(positive=['mr', 'industrial'], negative=['madam'], topn=1)
print(result)
```

```
[('ideas', 0.9467995166778564)]
```

Industrial + woman - man -> non-profit

Industrial + mr - woman -> ideas

This kind of biasing based on genders, occupation or any other factors where addresses in the paper and an approach to mitigate them was suggested.

Word-Scale Graphics:

Another research paper addresses the issues in word-scale graphics and visualizations embedded in text documents. It focuses on visualizing text data with respect to time or trends of a text document.

As given in the paper, one of the solutions to visualize word-scale graphics is SparkClouds which is an integration of tag cloud (word cloud) with sparklines. SparkCloud gives the how popular the word was over a definite period. ParallelCloud is another way of visualizing text data to represent the trends of words.