Sentiment Analysis on Twitter Data

FIT5147 - Data Exploration and Visualisation

Case Study:
Elections in India (Lokh-Sabha Elections - 2019)

Roopak Thiyyathuparambil Jayachandran - 28 April 2019
Student Id : 29567467
Tutor : Joy Zhao

Table of Content

Introduction

Democracy is understood to be a political system which affords voice to every participating citizen. Hence, every democratic exercise is formulated to efficiently transmit citizens opinion without misrepresentations. Lok Sabha election is one such democratic exercise, where in millions of citizens cast votes to influence their collective political future. Being the largest democracy in the world Lok Sabha elections demands unprecedented importance. Lok Sabha elections conducted in every five years to elect the lower house members know as Member of Parliament (MP). Candidates contest in elections from their respective constituencies and every adult citizen can cast vote in their constituency.

The 17th Lok Sabha elections are being conducted in India from 11 April 2019 to 19 May 2019. Currently ruling NDA government and the UPA headed by Indian National Congress are the main competitors in the 2019 elections.

Relevance of the topic

For sentiment analysis on Twitter data, it is important to select a topic which has highly polar opinions and many people supporting each side. Lokh Sabha(General) Elections 2019, India is one such topic which was trending #1 in Twitter for many days and since it is still in progress, people are still actively talking about it. Objective of the Exploration Project is to identify the sentiments of Twitter Users for each of the Party and check how these sentiments are related to different regions and keywords. This project will identify the online presence and impressions of each party per 1000 tweets and can give a rough idea on the strength and weakness of the political party.

Data Extraction

The main source of the data is Twitter. Twitter provides API for developers which can be used to extract data in Json Format. For this Exploration project **Tweepy** package in python is used to extract the data. Extracted data is then filtered to obtain just the required fields and other fields for explorations are generated through manual and natural language processing techniques.
In order to extract data from twitter, application has to be created in the official site which will generate consumer key, consumer secret key, access key, access secret key. These keys are needed which extracting the data. OAuthHandler and Cursor function from Tweepy are used with appropriate parameters. In the Cursor function the "keywords" and "number of tweets " is also specified. Twitter provides limitations to the extraction of large number of data for free but otherwise it is free of cost.

Code snippet which handles the same: (Python)

```python
auth = tweepy.OAuthHandler(consumer_key=consKey, consumer_secret=consSecret)
auth.set_access_token(_accessKey,  accessSecret)
api = tweepy.API(auth, wait_on_rate_limit=True)
tweets = tweepy.Cursor(api.search, q="BJP", lang="en").items(1000)
```

Extracted data is in Json format. Required fields are identified and data wrongling is done to convert semi structured data into structured data format.
Sample Json which was extracted:

```
'source': '<a href="http://twitter.com/download/android" rel="nofo
'in_reply_to_status_id': None,
'in_reply_to_status_id_str': None,
'in_reply_to_user_id': None,
'in_reply_to_user_id_str': None,
'in_reply_to_screen_name': None,
'user': {
  'id': 719401332893855744,
  'id_str': '719401332893855744',
  'name': 'pranshumishra',
  'screen_name': 'pranshumisraa',
  'location': 'Lucknow, India',
  'description': 'Journalist #India #UttarPradesh .Bureau Chief @C
  'url': 'https://t.co/xGvIjASrQQ',
  'entities': {
    'url': {
      'urls': [
        {
          'url': 'https://t.co/xGvIjASrQQ',
          'expanded_url': 'http://www.news18.com',
          'display_url': 'news18.com',
```

Platforms and packages used:

Stages	Language	Packages
Data Extraction	Python	Tweepy
Data Wrangling	Python	TextBlob, Pandas
Exploration and Visualisation	Python	WordCloud
	R	Leaflet, GGplot,
	Tableau	Bubble Chart, Bar Chart

Data Wrangling

Since the exploration involves Sentiment Analysis, **Textblob** python package is also used which will be used to find the polarity of the sentence. Polarity is then used to create buckets for each sentiment. In this exploration we have considered 5 Sentiments - Highly Negative, Negative, Neutral, Positive and Highly Positive.For example, if the polarity is identified between -1 to -0.5 then it is considered as a highly negative tweet.

```
created_date.append(each.created_at)
text.append(each.text)
longi.append(extractCoordinates(str(each.place))[0])
lati.append(extractCoordinates(str(each.place))[1])
impressions.append(each.retweet_count + each.user.favourites_count)
followers.append(each.user.followers_count)
user_location.append(each.user.location)


emotion = TextBlob(each.text)
emo = emotion.sentiment.polarity
if emo >= 0.5:
    polarity.append("Highly Positive")
if emo <= -0.5:
    polarity.append("Highly Negative")
if emo == 0:
    polarity.append("Neutral")
if 0 < emo < 0.5:
    polarity.append("Positive")
if -0.5 < emo < 0:
    polarity.append("Negative")

pol_value.append(round(emo, 2))
```

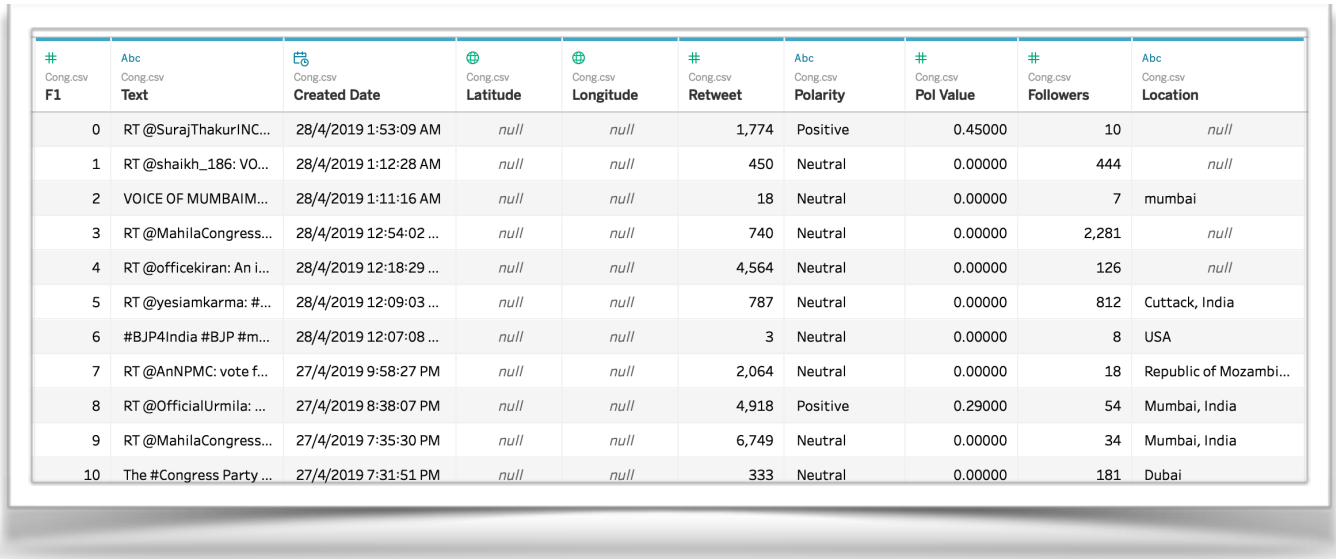
Similarly following assumptions have been made for identifying the buckets:

Polarity Value	Group
-.1 to -0.5	Highly Negative
-0.5 to 0.0	Negative
Exactly 0.0	Neutral
0 to 0.5	Positive
0.5 to 1	Highly Positive

Fields identified are : Tweet, created date, latitude, longitude, retweets, Polarity, polarity_value, followers and location. The fields are extracted from the JSON , inserted to a Dataframe and later saved to a csv file.

Data Checking

After completion of Data wrangling, 2 csv files are generated, one for each of the political party. On checking the csv files, it was identified that the coordinate fields are missing for most of the tweets. Coordinates are missing as Twitter wants to ensure the privacy of the twitter user. Coordinates are only available if the user has enabled the setting in privacy options. Sample of the data stored in csv file looks as below:

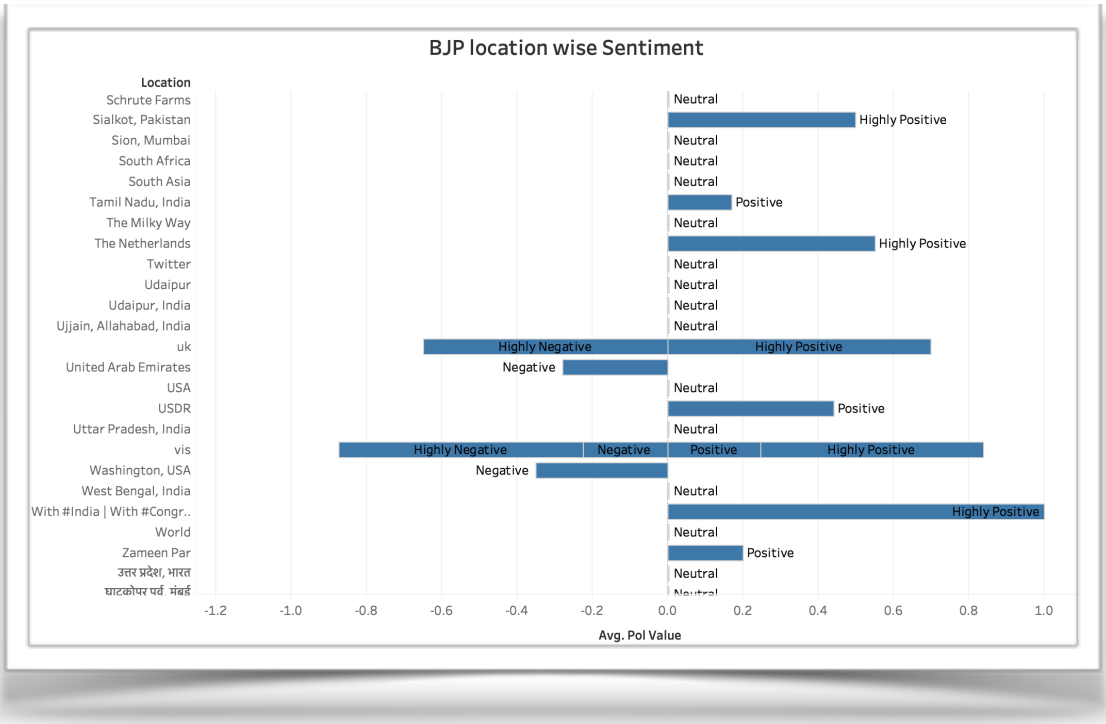


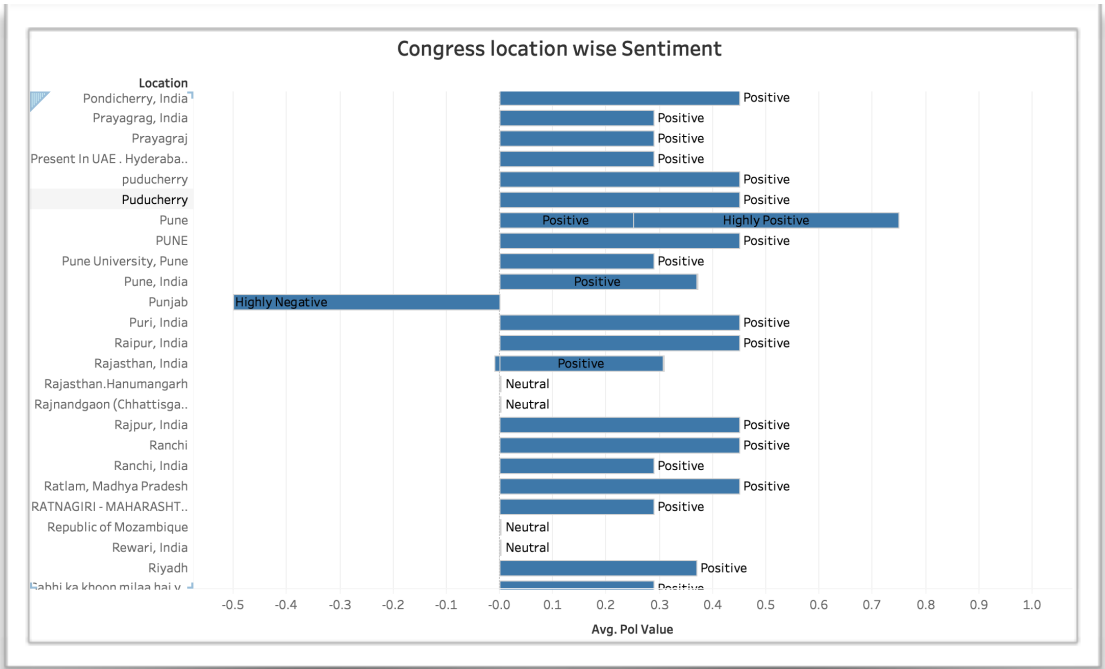
	F1	Text	Created Date	Latitude	Longitude	Retweet	Polarity	Pol Value	Followers	Location
	Cong.csv	Cong.csv	Cong.csv	Cong.csv	Cong.csv	Cong.csv	Cong.csv	Cong.csv	Cong.csv	Cong.csv
0	0	RT @SurajThakurINC...	28/4/2019 1:53:09 AM	null	null	1,774	Positive	0.45000	10	null
1	1	RT @shaikh_186: VO...	28/4/2019 1:12:28 AM	null	null	450	Neutral	0.00000	444	null
2	2	VOICE OF MUMBAIM...	28/4/2019 1:11:16 AM	null	null	18	Neutral	0.00000	7	mumbai
3	3	RT @MahilaCongress...	28/4/2019 12:54:02 ...	null	null	740	Neutral	0.00000	2,281	null
4	4	RT @officekiran: An i...	28/4/2019 12:18:29 ...	null	null	4,564	Neutral	0.00000	126	null
5	5	RT @yesiamkarma: #...	28/4/2019 12:09:03 ...	null	null	787	Neutral	0.00000	812	Cuttack, India
6	6	#BJP4India #BJP #m...	28/4/2019 12:07:08 ...	null	null	3	Neutral	0.00000	8	USA
7	7	RT @AnNPMC: vote f...	27/4/2019 9:58:27 PM	null	null	2,064	Neutral	0.00000	18	Republic of Mozambi...
8	8	RT @OfficialUrmila: ...	27/4/2019 8:38:07 PM	null	null	4,918	Positive	0.29000	54	Mumbai, India
9	9	RT @MahilaCongress...	27/4/2019 7:35:30 PM	null	null	6,749	Neutral	0.00000	34	Mumbai, India
10	10	The #Congress Party ...	27/4/2019 7:31:51 PM	null	null	333	Neutral	0.00000	181	Dubai

Explorations and Questions

At this point, data extraction is done using Twitter API which was in JSON format (unstructured) and stored in CSV file(structured) format with some wrangling necessary for visualization. Following are the question which will be answered using Exploration and Visualisation

1. Perception of parties as per geography

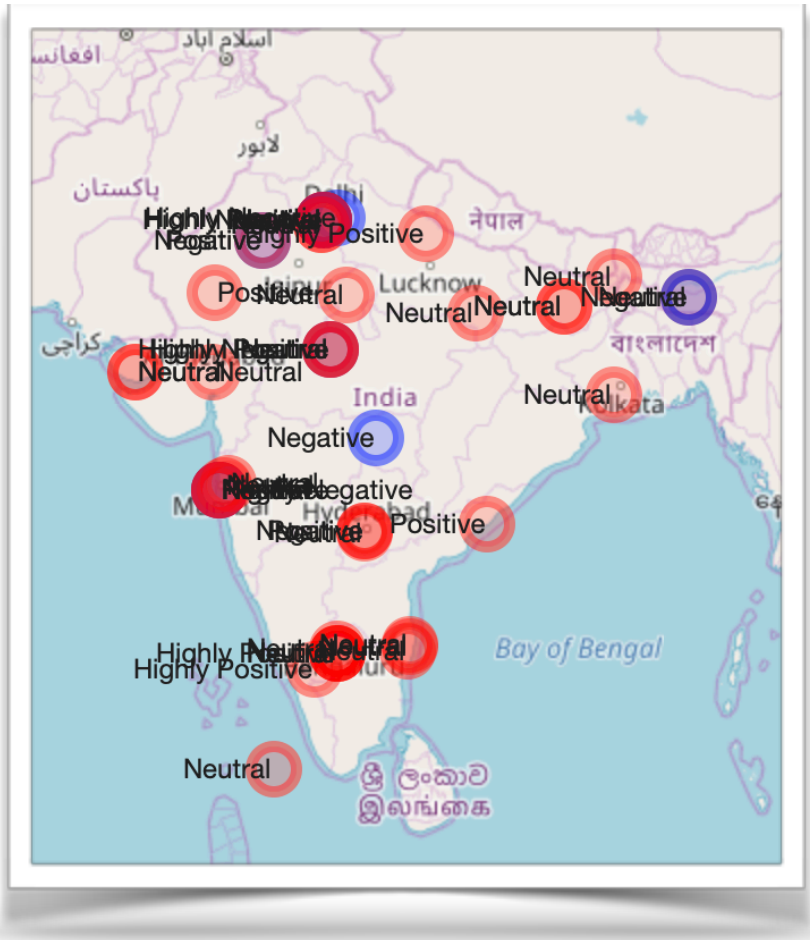




This exploration is done using Tableau in which we can see for each region, what is the general opinion of the

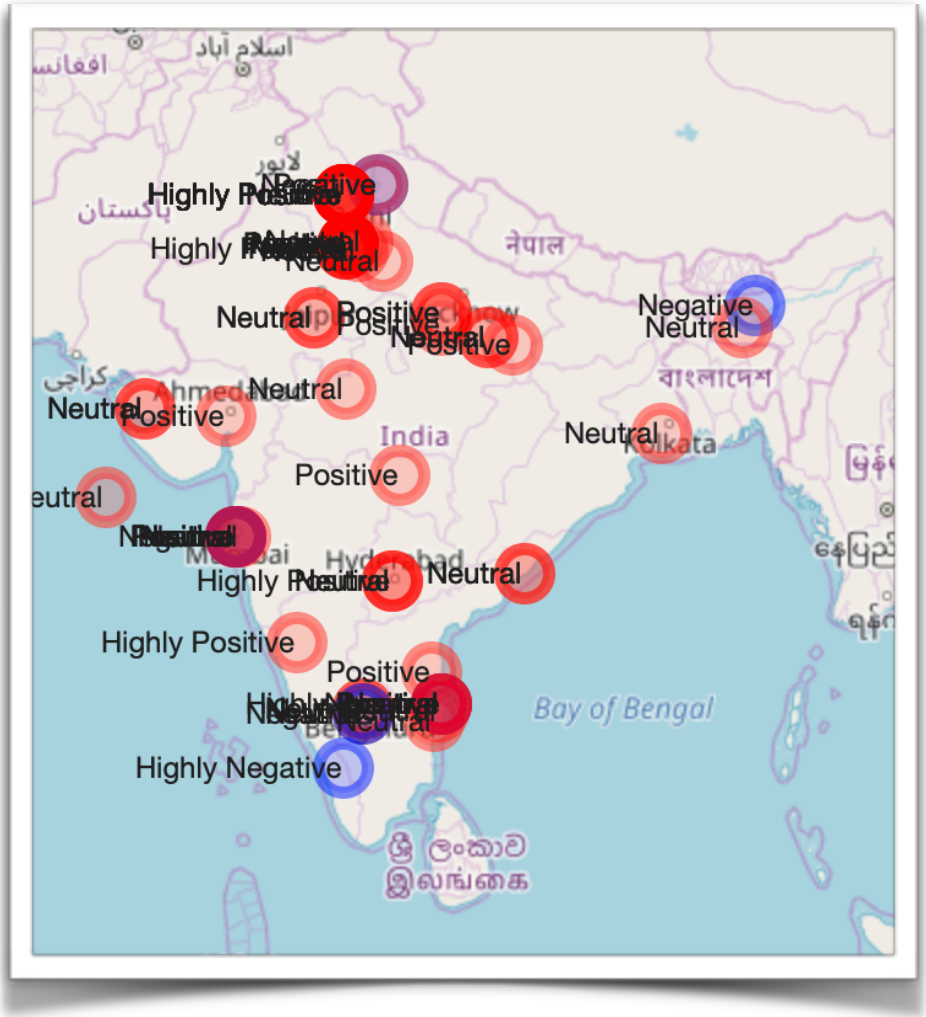
public(Twitter user). The attached screenshot is not a complete visualisation as there are many more regions which are identified from the Twitter data. But on average it can be noticed that people in most of the regions are tweeting more positively for Congress while in the given screenshots BJP has more neutral tweets to its name. Also the negative tweets are present more for BJP. Bur for Congress Punjab region shows a highly negative average of Sentiment.

Leaflet in R can be also used to do the visualisation for this problem. As the available coordinates are very less in number and also due to overlapping of sentiments in the similar coordinates, the visualisation gives very rough idea on peoples perception.



In this Visualisation, First one is of Congress and second one is of BJP. Blue represents negative and Highly negative tweets whereas Red represents positive, highly positive and neutral tweets.

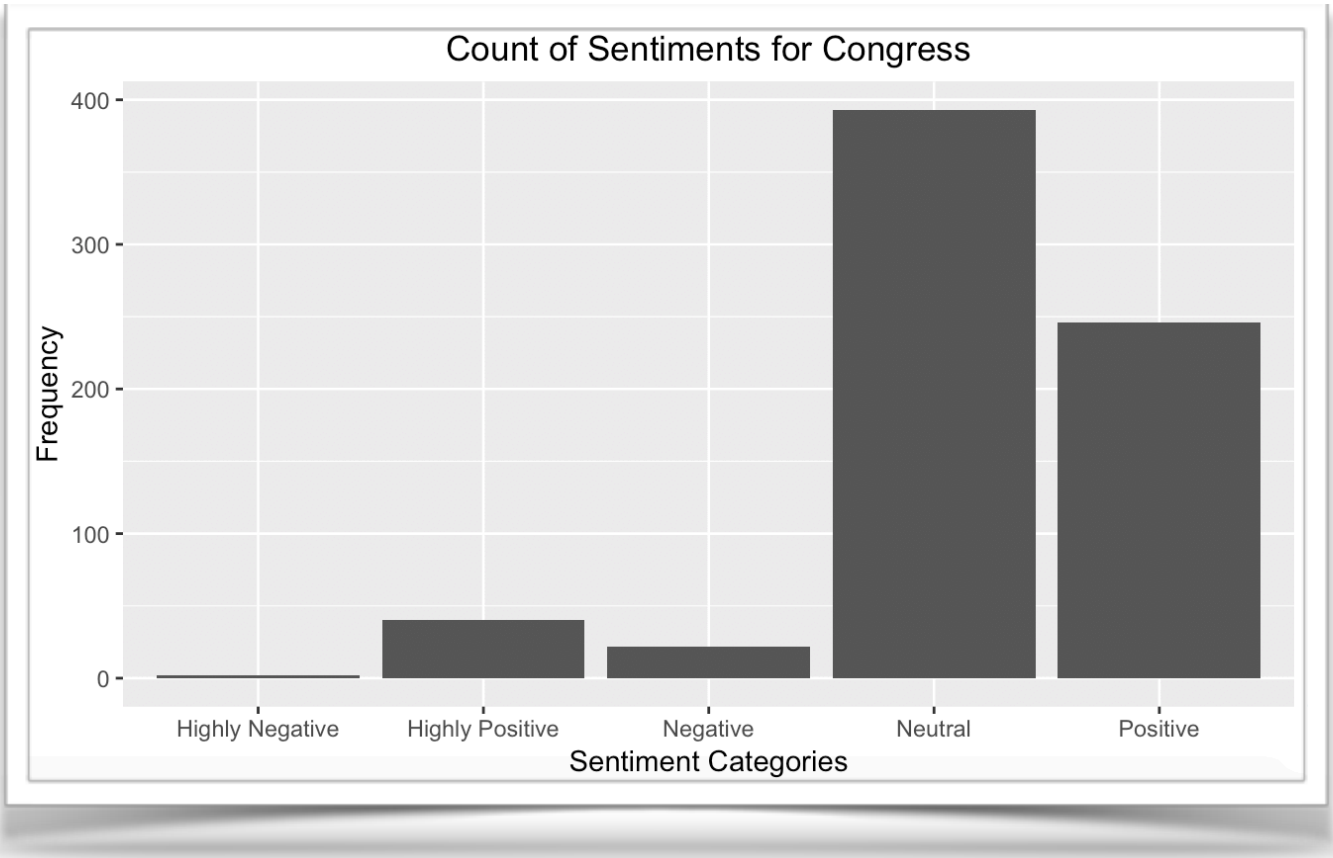
In the first plot, most of the tweets are neutral or positive where as some tweets are also negative towards the east. It can be also noticed that rate of positive tweets towards east is very less as most of the tweets are either negative to neutral.

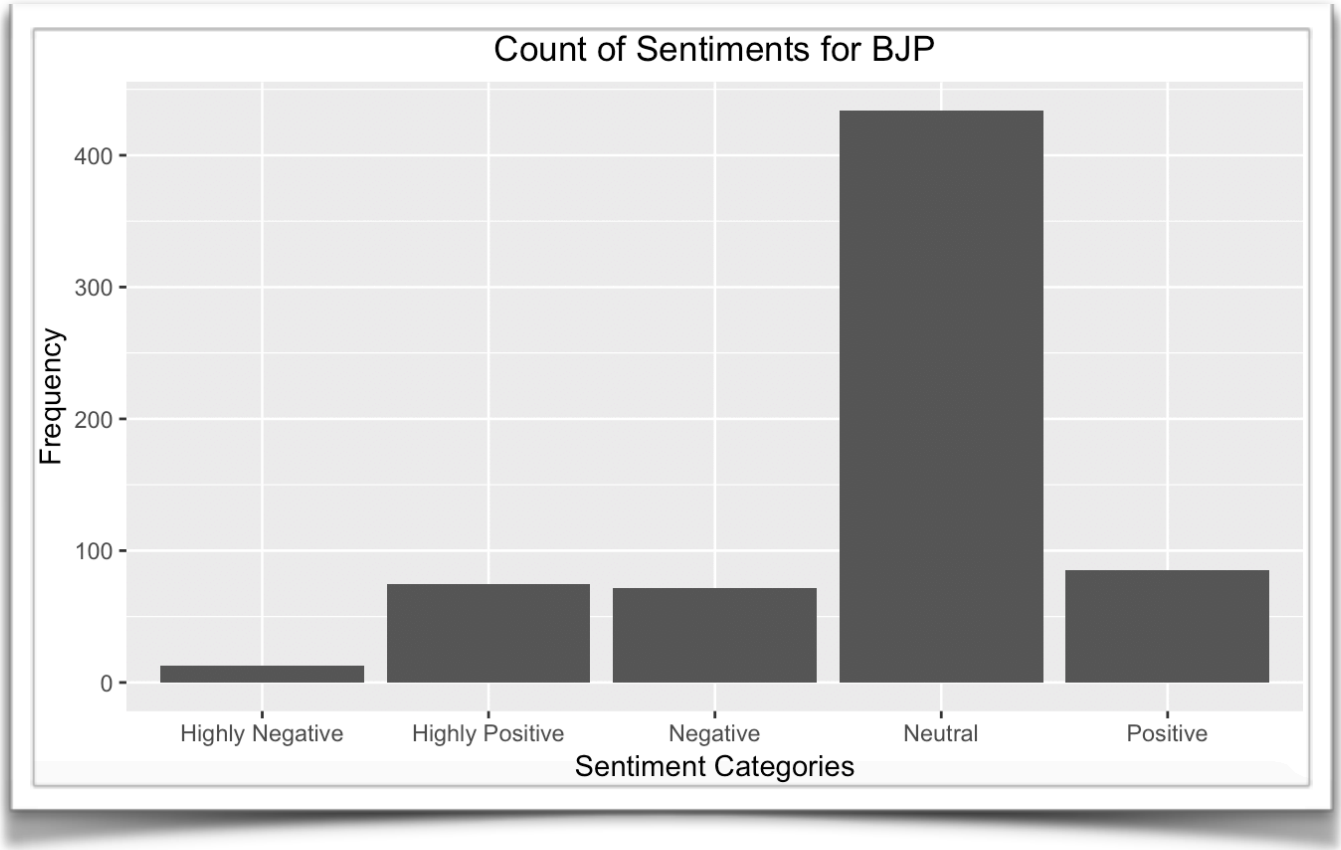


Here also blue represents negative and Highly negative tweets whereas Red represents positive, highly positive and neutral tweets.

Plot for BJP also show that most of the tweets are positive or neutral. But for BJP unlike Congress South India also has some "Highly Negative" tweets to its name.

2. What are the sentiments linked to each party



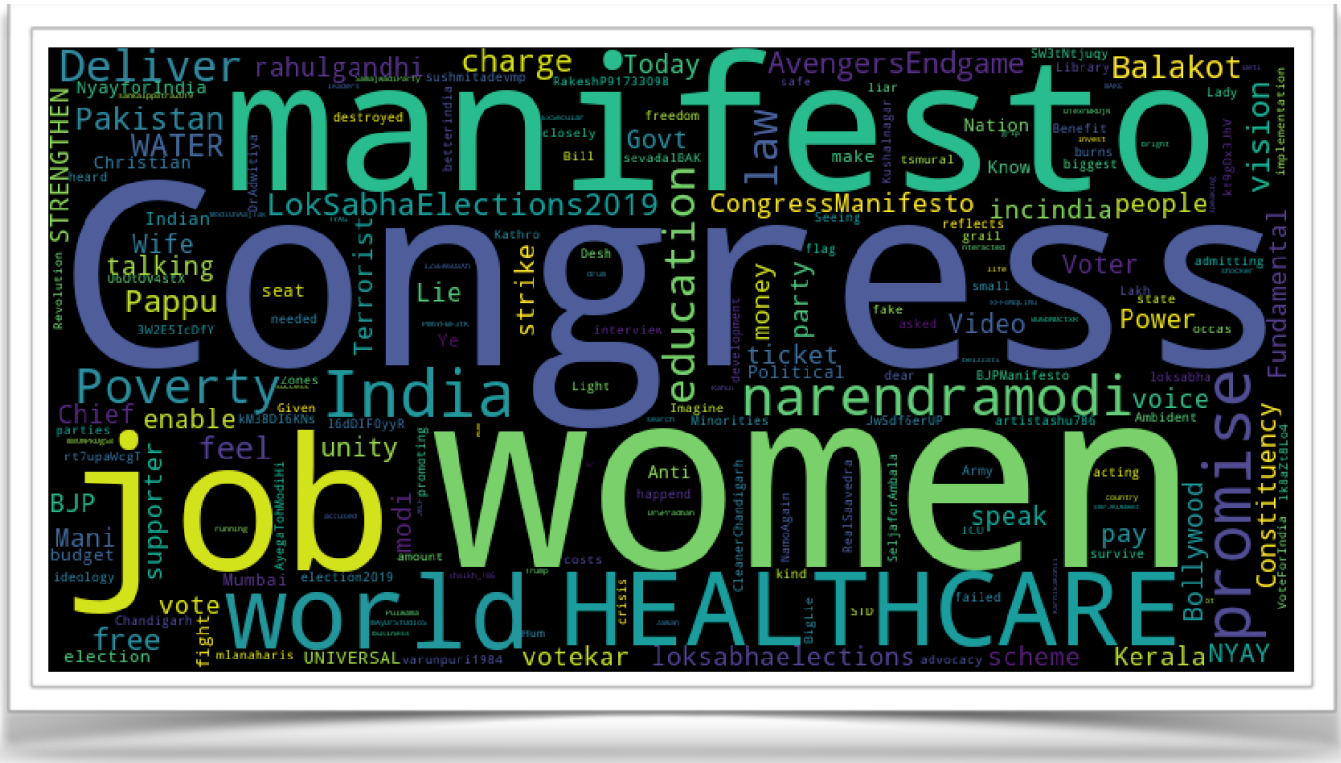


The above two plots in R, shows the count on Sentiments for each category in general per 1000 tweets. Congress and BJP both have a maximum for Neutral tweets while the number of Highly positive tweets are more for BJP. But along with this they also have an equal number of negative tweets. In the comparison, Congress has more Positive tweets which is more than double of BJP.

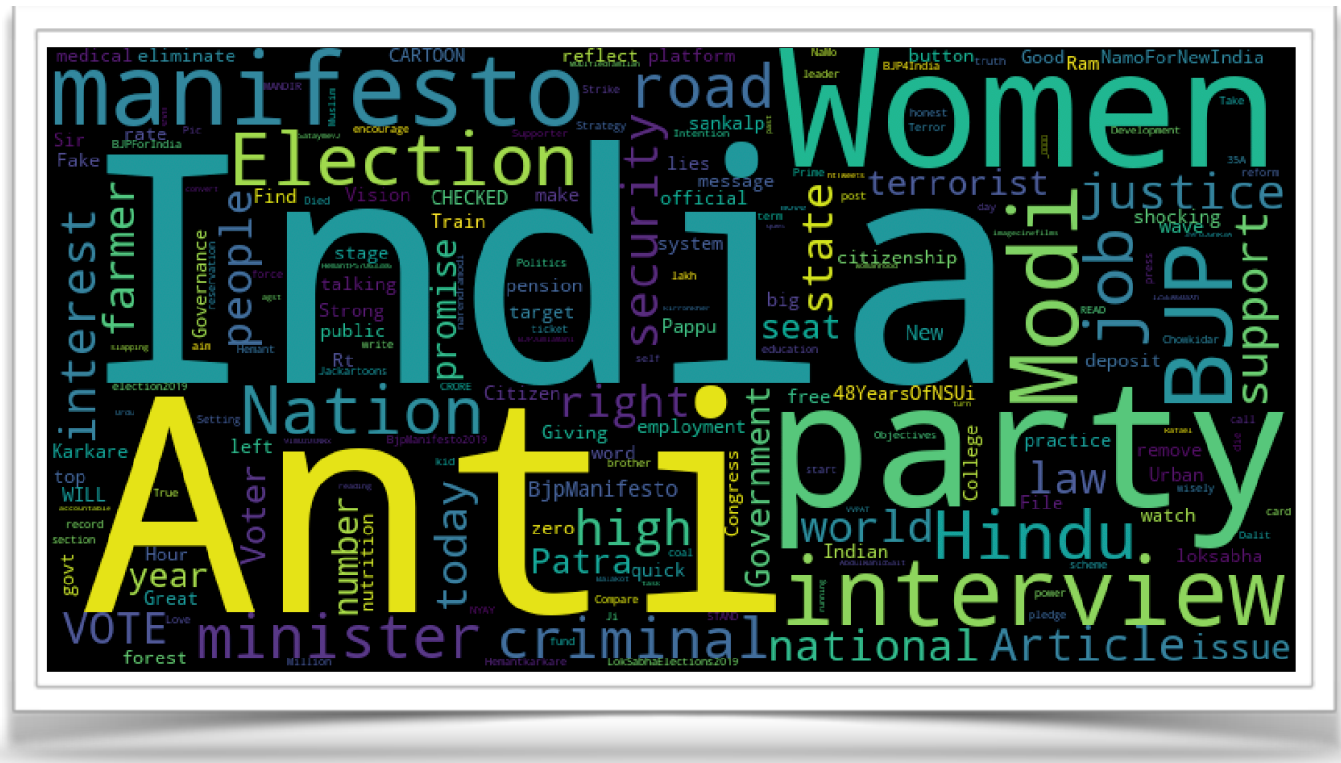
3. Focus points of Each party as identified by Users

Extracted tweets were converted into tokens and using Natural Language Processing nltk python packages, top keywords were identified. For this process stop words were also removed and a keyword vocabulary was formed for each party. WordCloud representation using Python WordCloud package can be used to find what keywords Twitter users have identified which was focussed by each of the party.

As it can be noticed, Users have identified that Congress focuses more on women empowerment, jobs, poverty reduction, healthcare etc

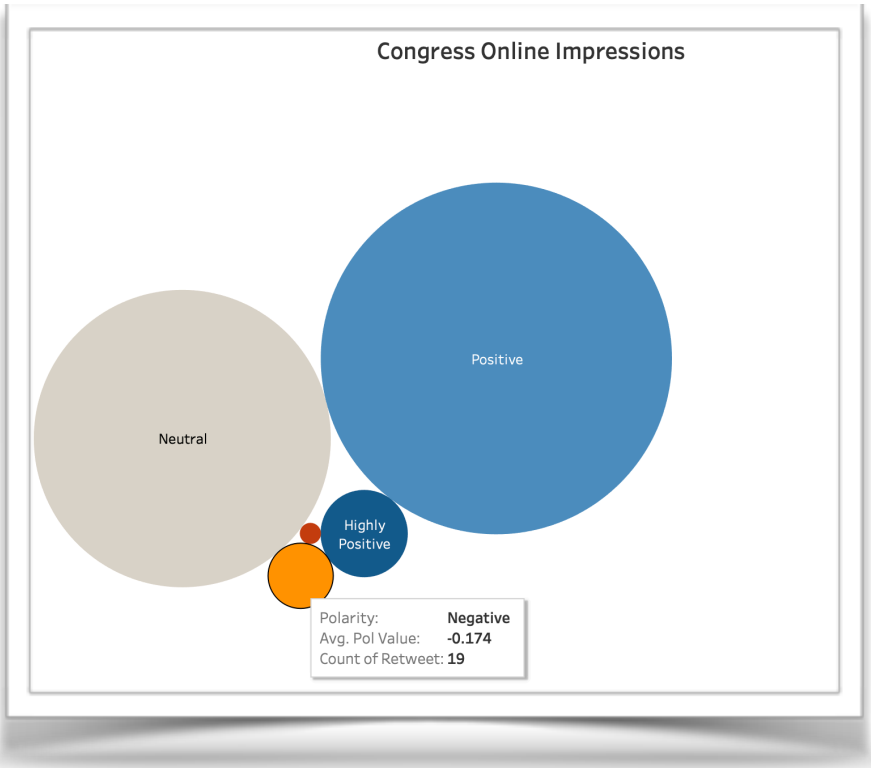


Whereas BJP concentrates on Hindutwa, Jobs, Women empowerment, New and modern India etc



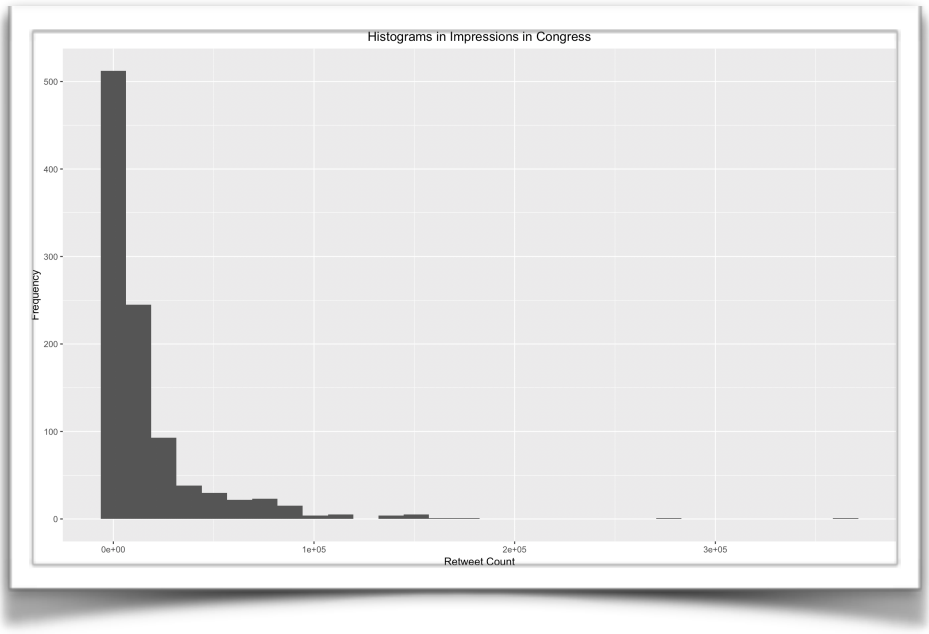
4. Online impression for each party

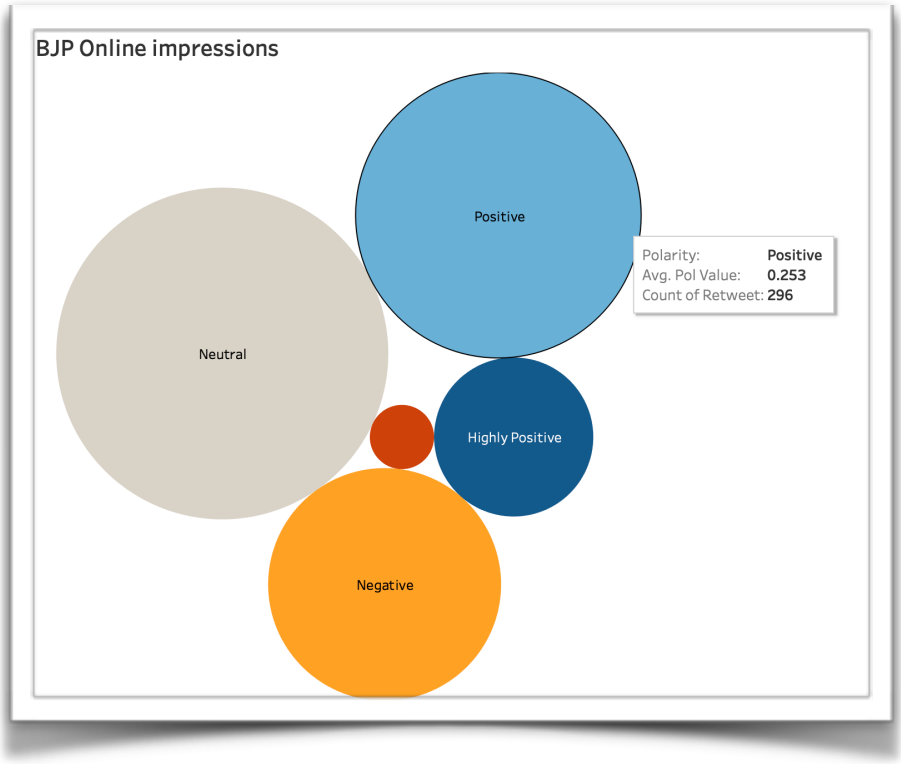
Retweets, favourites and shares are effective measure of Twitter engagement by determining the interests and preferences among the users. It was researched that tweets with one retweet get at least one more retweet 43% of the time. This ensures more reach towards public and can help in the public relations of the Party. On Twitter, metrics such as favourites speak volumes to engagement with followers. In the column "retweet" we have combined both retweet and favourites count so that the overall "twitter impressions" can be identified.



The bubble chart shows the quantity of Positive, Negative, Highly Positive, Highly Negative and Neutral retweets that Congress has in 1000 tweets. As it is seen Positive and Negative retweets are very high in number where as others are less

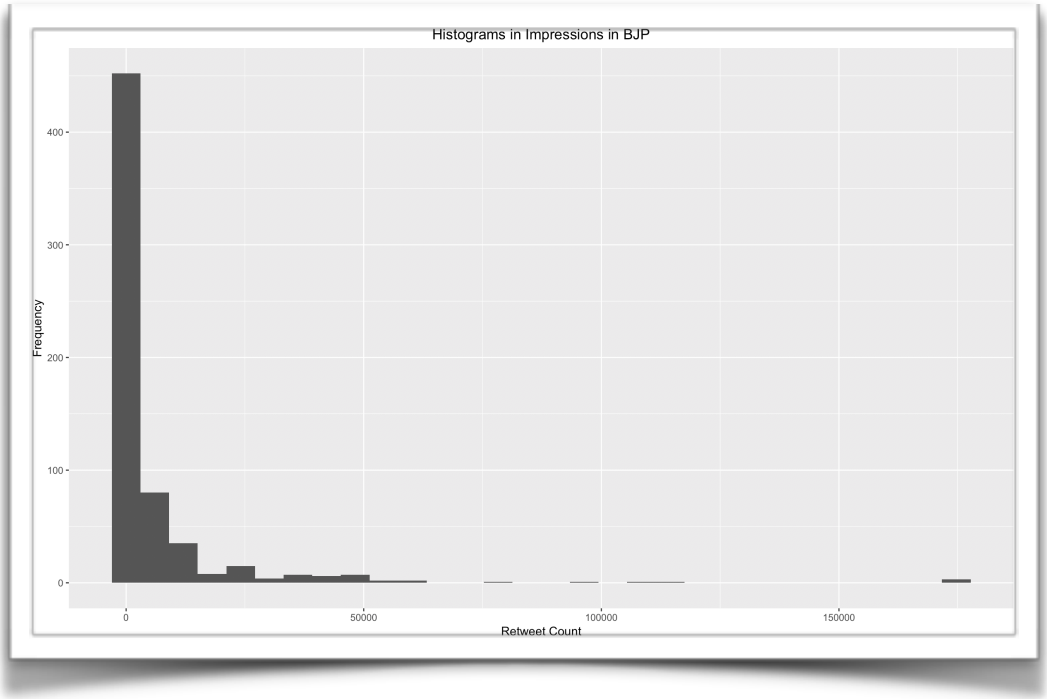
The second plot represents the histogram of twitter impressions in general per 1000 tweets. It can be noticed that most of the tweets are not retweeted and there are very few tweets which have very high number of retweets





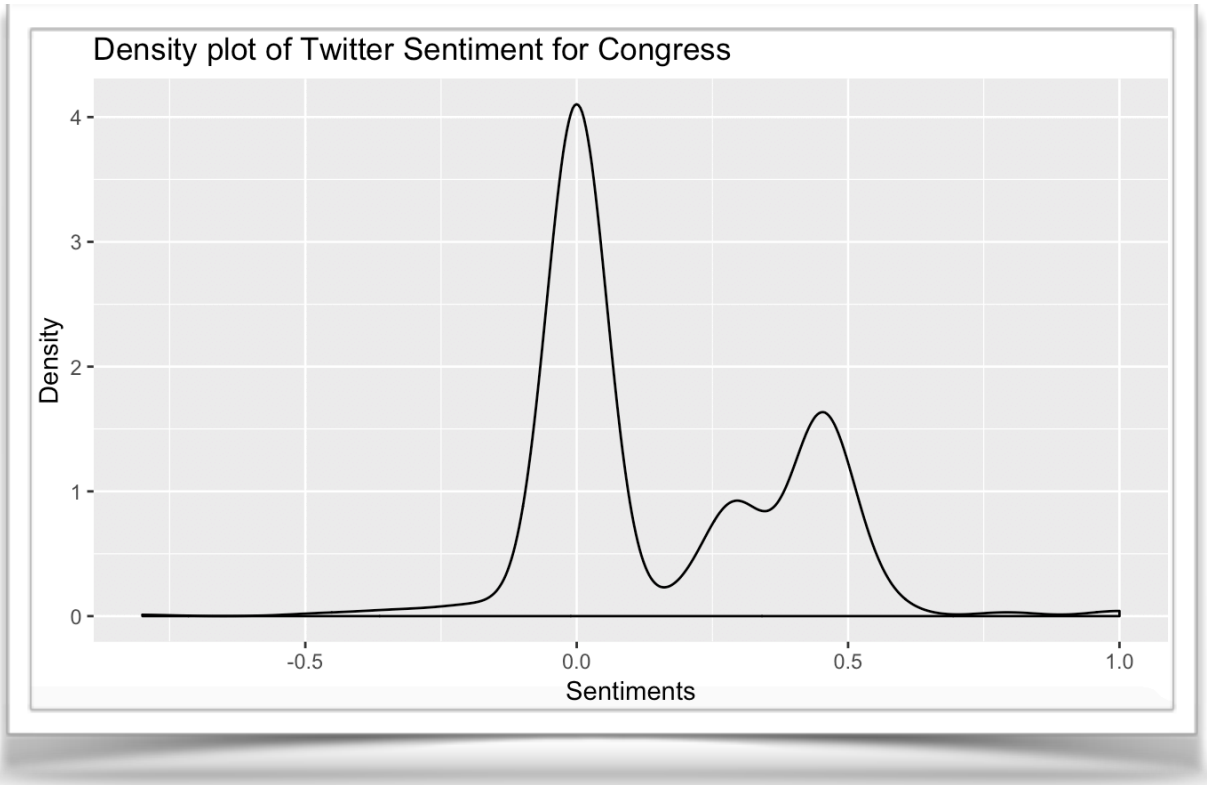
On the other hand for BJP Positive, Negative and Neutral tweets have been retweeted the most and also it has more "Highly Positive" retweets that Congress. It concludes that per 1000 tweets BJP has a better online impressions are online PR firms are working well for them

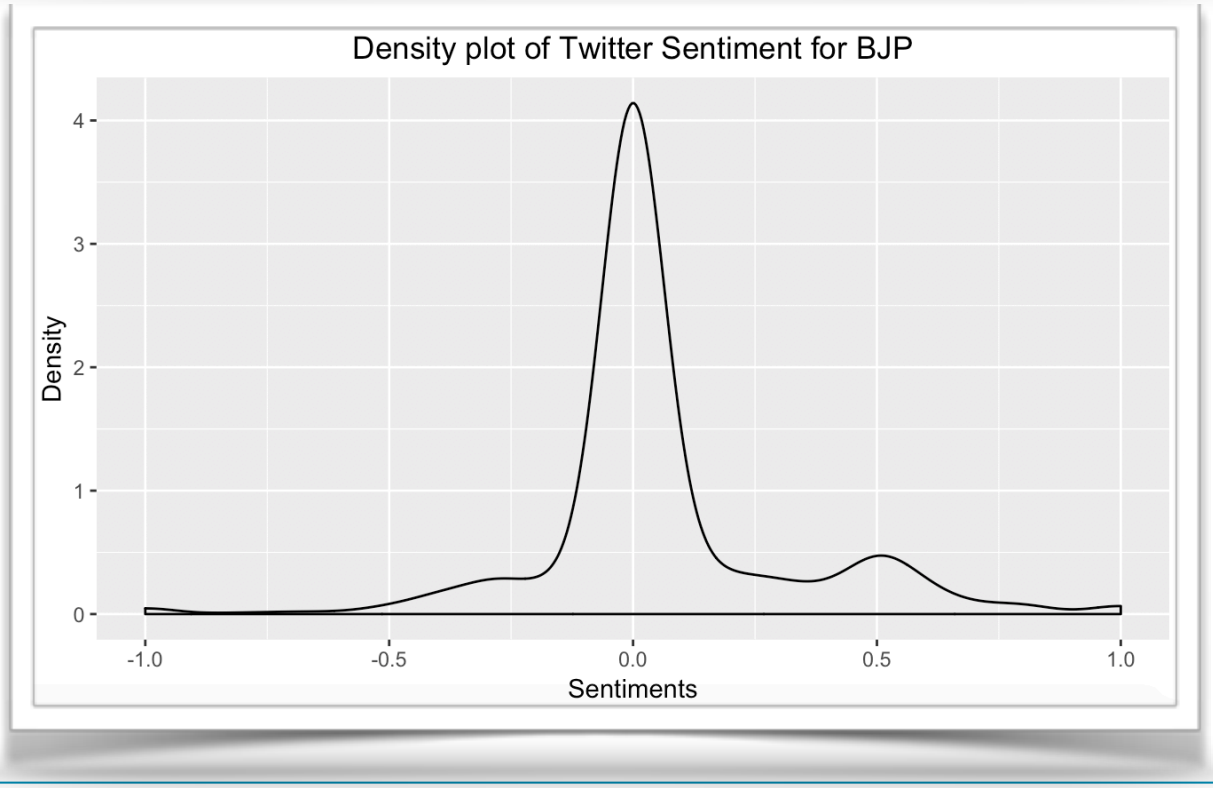
For general retweets, BJP has more tweets than congress per 1000 tweets which were not retweeted but the number of tweets with huge retweets are more for BJP



5. Distribution of the Sentiment

On plotting both the distribution of sentiment using R (GGplot), it can be clearly identified that both the plot similar distributions but Congress has more positive tweets





Conclusion and Limitation

- For this Exploration only a sample of Population is considered, I.e., 1000 tweets for each Political party. This sample can not be used to identify or predict the original chances of winning as millions of tweets are generated each day on these hashtags and less than 50% voting population of India is active on Twitter to cast their opinion.
-
- This Exploration can just give a very rough idea on what all are the online Twitter population is talking about, are they happy or sad for any step or action which the government has taken. What are the key points which are in favour and against the party and in which all places, people have a very positive opinion on the party. This can help the party to concentrate on the areas for which people are not happy about which can help in improving the chances of winning the elections.
-
- For actual prediction of elections a dedicated team has to be employed which will do the analytics in real-time continuously and do the analysis. That report will thus by utilised by the party and the PR (Public Relations) firms which are associated with the party during the elections to plan accordingly

Reflection

In this Exploration and Visualisation project, different platforms are packages were used for extraction, wrangling, exploration and visualisation. There was a big learning curve to use these packages simultaneously and come up with results. The whole project can be more improved by considering tweets from the previous election time period and using that to compare with the current scenario. Using more more tweets can also be beneficial. Also with more tweets with coordinates available, region-wise exploration could have been improved.

References

[1] TJ, R. (2019). Docs. Retrieved from https://developer.twitter.com/en/docs.html
[2] TJ, R. (2019). The R Graph Gallery. Retrieved from https://www.r-graph-gallery.com/
[3] TJ, R. (2019). Gallery. Retrieved from https://public.tableau.com/en-us/s/gallery