COSC2671: Assignment 2

Twitter data analysis to find the most influential Minister in India

[Vijeta Tulsiyan – s3398979

(s3398979@student.rmit.edu.au)

Salina Bharthu – s3736867

(s3736867@student.rmit.edu.au)]

COSC 2671 | Social Media and Network Analysis
Date: 13-10-2019

Contents

Introduction

In recent years, social media has become widely used platform for politicians in-order to communicate with people about various current affairs faced by country or election campaign. As it bypasses any editorial media, people tend to prefer social media as a source of information. Therefore, it has become open platform for interacting with the elected ones, debating and critiquing instantaneously. By collecting, analyzing, summarizing , and visualizing politically relevant information from social media, the influence of political party or leader can be inferred. As India is world's largest democracy with the population of over 1.3 billion people, having 8 recognized national parties and year 2019 has been very important for Indian politics, as the general elections were held after 5 years, many interesting inferences can be gathered.

This study aims at finding the most influential political figure in India and study the social network of political figure. It involves analyzing data which refers to Indian politics and current affairs and finding out top two talked about political figures of Indian politics, followed by comparing of both politicians in-order to find out the most influential one.

Executive Summary

We analyzed the twitter data to find the two most influential leader in politics in India. We found Narendra Modi and Rahul Gandhi are most talked about political figures. We analyzed the tweets from both the leaders and compared what keywords they tweet, their global presence, tweets' sentiment (positive / negative) and number of followers and followed. We found that Narendra Modi is more influential compared to Rahul Gandhi in terms of a large global presence of its followers, most positive keywords tweets, stable positive sentiment and large number of followers.

Data Collection & Preparation

For initial analysis to find popular political figures, data was extracted from reliable news account twitter handle "TimesNow". For further comparison, sentiment analysis of those popular political figures (Narendra Modi and Rahul Gandhi), data is collected using their Twitter handle and hashtags separately. All data is fetched using rest API.

Similarly, for event detection, data is collected using the twitter handle for Narendra Modi and Rahul Gandhi separately. Around 5000 tweets are extracted over a period of one year using rest API.

The noisy tweet data collected at every step is pre-processed using below mentioned steps:

- Tokenization: Tokenized the data using tweet tokenizer of nltk library to convert each tweet into tokens of words.
- Stop word removal: Removed punctuations and stop words (such as '"', 'ji', 'day', '…', 'see') which are not useful for analysis.
- Used regular expressions to remove emojis, digits and urls from data.

For Social Network Analysis, due to twitter API limitation and efforts to create a visual appealing graph, limited data is used for graph. The API allowed only few hundreds of followers tweets to be extracted, whereas both the leaders have followers in millions. We worked a ratio of followers and followed count so that they can be easily visually presented in a graph.

Data is prepared using the text preprocessing steps such as tokenizing, stemming, stopwords removal. Python NLTK library is used for data cleaning.

Data Exploration

In this step, we are exploring and analyzing twitter data:

- ### Finding Most frequent keywords:

1. Fetching Twitter data related to Indian politics:

To find highly talked-about politicians, we have fetched twitter data of the TimesNow news account. **TimesNow** tweets about current affairs and trends of Indian politics. We have fetched 3k tweets from @TimesNow twitter handle.

After initial analysis and pre-processing of the tweets, plotted wordcloud of frequently occurring terms in tweet data.

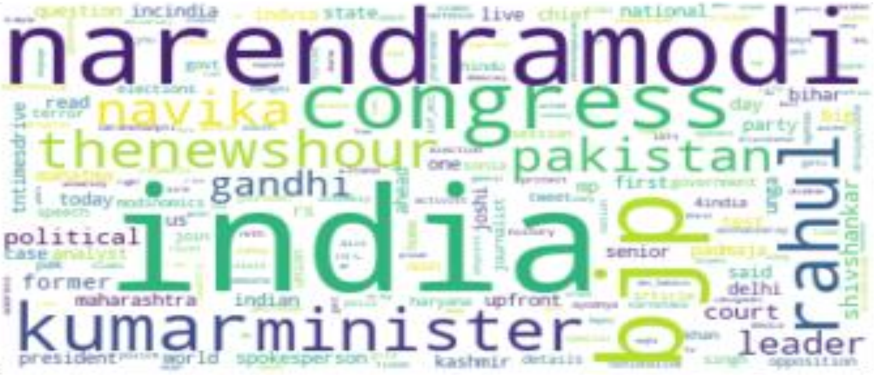


Figure 1- Word Cloud of Frequently mentioned words in TimesNow Tweets

From the above wordcloud, we can see frequent words such as narendramodi, India, bjp, congress, Rahul Gandhi, minister, Pakistan, etc., which shows the topics of frequent discussion over the news.

It can be inferred that, after India, Narendra Modi is mentioned most frequently along with the name of political party “BJP”, which he is leading at present. Moreover, the name of Rahul Gandhi is also mentioned multiple times in tweets along with the name of political party “Congress”, which is led by him currently.

As “BJP” and “Congress” is the top two political party of India and we can see that the leaders of both parties are discussed the most, we have performed our further analysis on the twitter data of Narendra Modi and Rahul Gandhi.

2. Comparing most frequently used words by both politicians:

After finding Narendra Modi as a most addressed politician over the news, we have fetched the twitter data of him to find inferences about him. The analysis is performed over 5k tweets extracted using twitter handle ‘@narendramodi’.

Similarly, twitter data of Rahul Gandhi is extracted. The initial pre-processing is performed including tokenization and stop word removal on both datasets.

Now, to find out the most frequent term both politicians use in their tweets, the pre-processed data is used. Below word clouds depict the same:

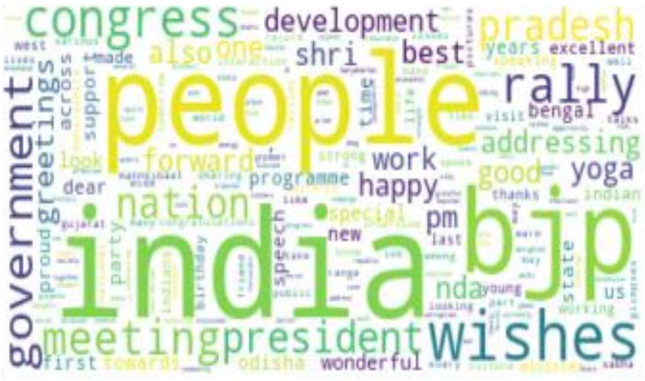


Figure 2- Frequently used words by Narendra Modi

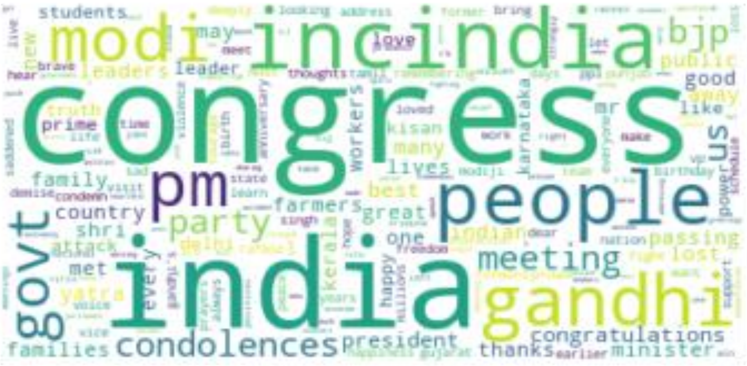


Figure 3- Frequently used words by Rahul Gandhi

As shown in the word cloud of Narendra Modi, he uses India, people and BJP(the political party led by him) the most. However, he also talks about opposition party congress, there is no direct mention of the leader of the party, Rahul Gandhi.

Also, it can be seen from the word cloud of Rahul Gandhi that, he uses words like India, People, Congress and INCIndia(the political party led by him) the most. Moreover, he mentions Modi and bjp as well in many tweets.

- **Comparing Hashtags Count:**

In this step, we are fetching top 15 hashtags which are occurring the most in the tweets of Narendra Modi and Rahul Gandhi separately.

Below bar plot shows the hashtags with its frequency:

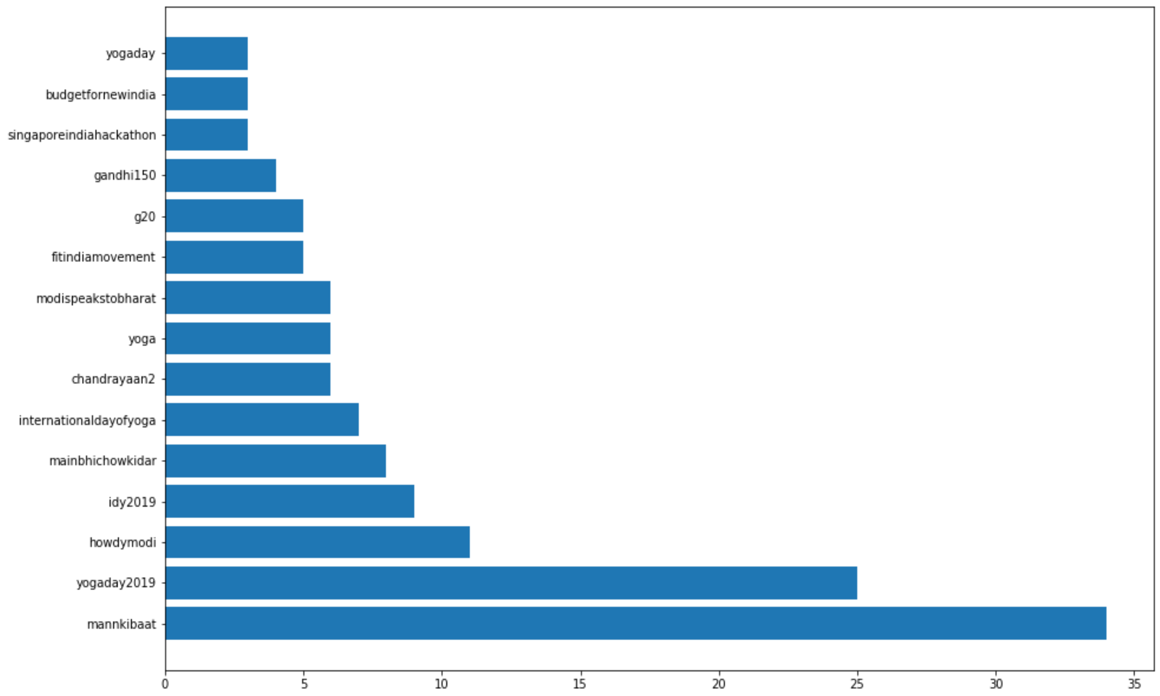


Figure 4- Frequently used hashtags by Narendra Modi

From the above plot, it can be inferred that, ‘mannkibaat’ (Heart’s voice) is the most occurring hashtag which refers to the Indian program hosted by Narendra Modi. Using this platform, he interacts the people of nation. After this, ‘yoaday2019’ is occurring the most. There are other hashtags related to Yoga as well, which shows his inclination to spread awareness about fitness and Yoga among people. Followed by Yoga Day hashtag, ‘Howdymodi’ hashtag is used the most, which refers to the latest event hosted for Narendra Modi and the president of US Donald Trump Houston, Texas.

Below plot depicts the most frequently occurred hashtags in Rahul Gandhi’s tweets.

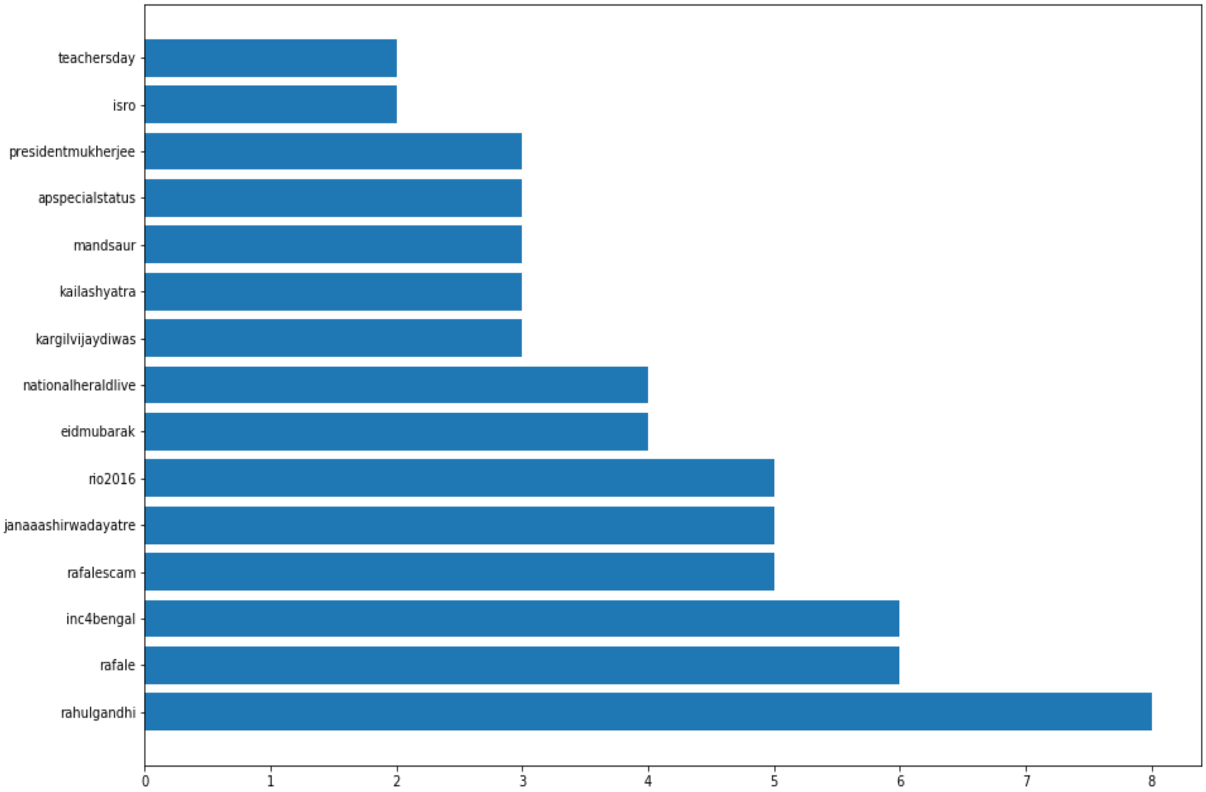


Figure 5- Frequently used hashtags by Rahul Gandhi

Above plot shows that, “rafale” is the most occurred hashtags after the hashtag with his own name. This hashtag refers to the fighter aircrafts bought by Defense ministry of India. After this, “INC4bengal” hashtag is appearing the most, which refers to the campaigning by his political party for West Bengal elections.

- ## Comparing retweet frequency of their tweets:

To further compare the influence of both politicians, we have fetched the retweet count of their tweets and plotted it over the time period of 1 year.

Below time series plot shows the retweet frequency:

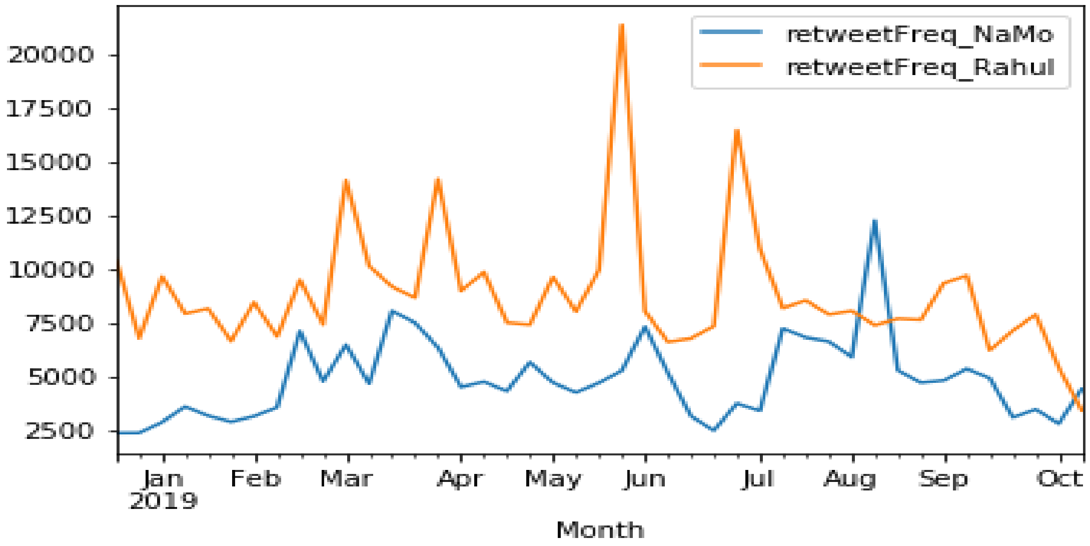


Figure 6- Plot comparing retweet frequency of tweets

Above plot depicts that, tweets of Rahul Gandhi are tweeted a greater number of times as compared to the tweets of Narendra Modi.

Now, to understand this influence with regards to positive or negative influence, we have performed sentiment analysis on their tweets.

- **Finding the presence across the Globe:**

  In this step, to further understand popularity of both political figures, we have fetched the twitter data of '#narendramodi' and '#rahulgandhi'. The data is pre-processed, and location of user is fetched from the tweets. The location coordinates fetched from the tweets related to Narendra Modi, are plotted on the below map.

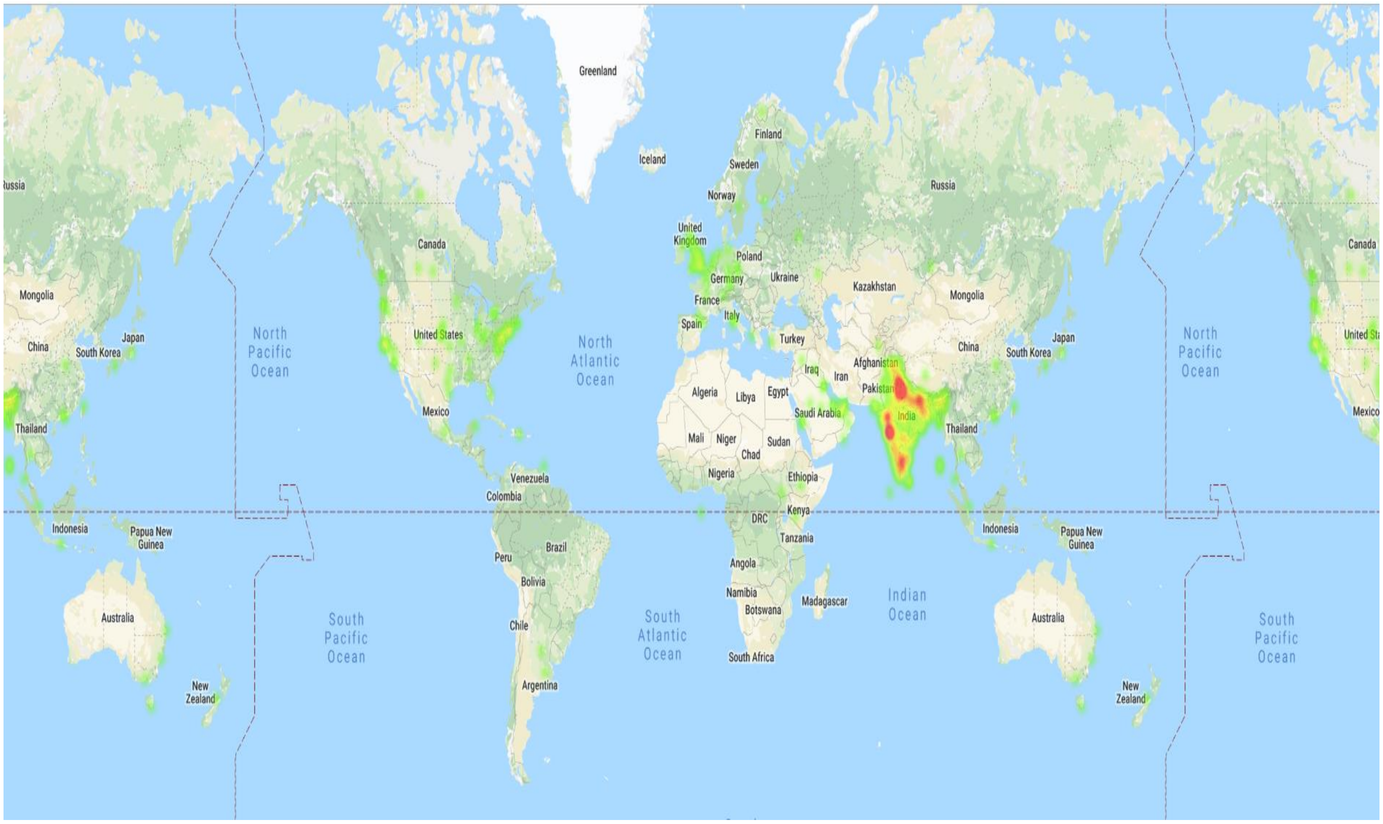


Figure 7- Map of location where Narendra Modi is mentioned in tweets

From the above map, we can infer from the intensity of color that, India is having highest number of people mentioning Narendra Modi in their tweets. Apart from India, people from Pakistan, UK, Germany, France, USA and Australia also mention him in their tweets. This shows that Narendra Modi is widely discussed around the globe.

Now, to further analyze the location data, we have fetched the location tag data from the tweets. The location data is first pre-processed in-order to remove trailing whitespaces and redundant location names.

After pre-processing, we have plotted the bar chart of top 35 locations of people mentioning Narendra Modi and Rahul Gandhi separately.

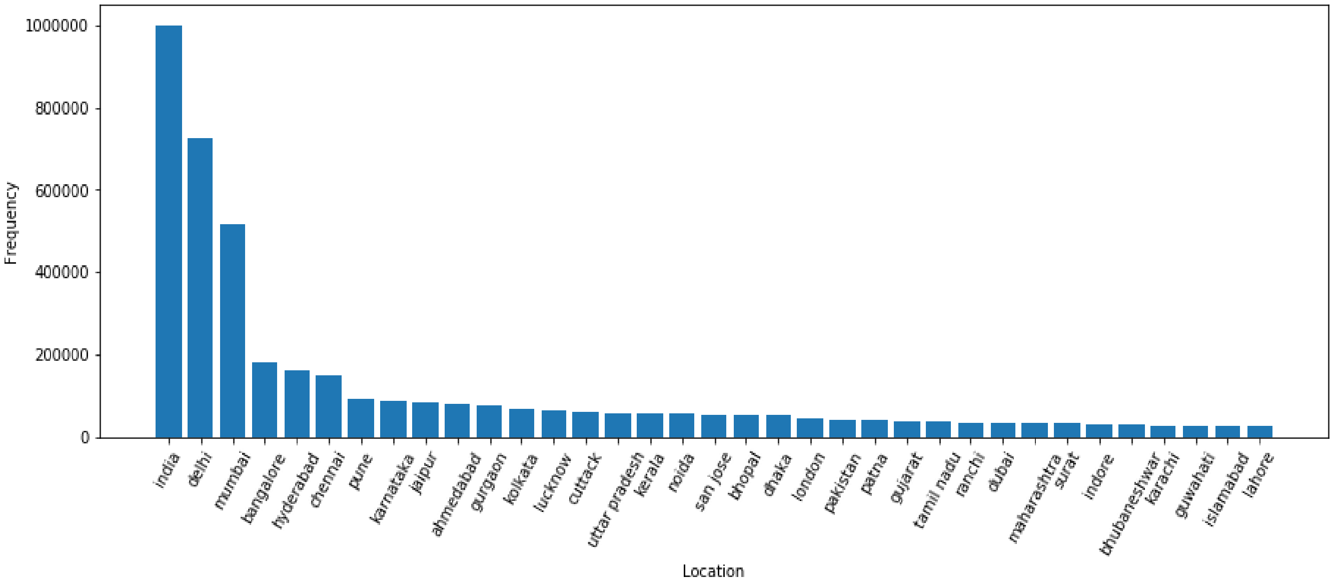


Figure 8- Location where people talk more about Narendra Modi

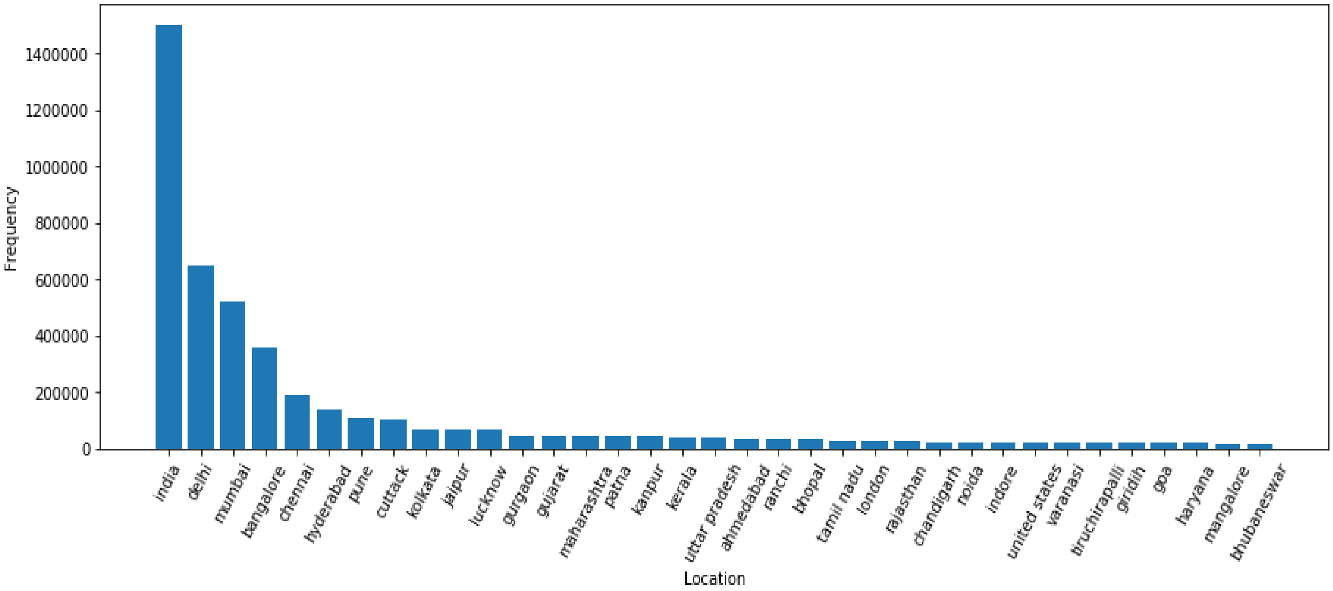


Figure 9- Location where people talk more about Rahul Gandhi

From above bar plots, it can be inferred that people from the cities such as Delhi, Mumbai, Hyderabad and Chennai are having higher number of people talking about both Narendra Modi and Rahul Gandhi. The top 35 location of Narendra Modi mentions beholds locations such as san jose, Landon, Pakistan, etc. depicting global presence of Narendra Modi. Similarly, there are locations such as united states, London, etc., depicted in the bar plot of Rahul Gandhi, which is comparatively less widespread than

Narendra Modi. Hence, to further understand their impact, we have implemented Sentiment analysis, Event detection and Social network analysis.

Sentiment analysis

To perform the sentiment analysis of tweets of Narendra Modi and Rahul Gandhi, the pre-processed tweet data is used.

- **Vader Approach:**
  As Vader approach of sentiment analysis provides many features such as punctuation, capitalization, degree modifier and polarity, and generates the sentiment scores considering the semantic orientation of each word, we have chosen Vader approach.

  After generating compound sentiment score for tweet data, we have merged the results of both Dataframe (Narendra Modi and Rahul Gandhi tweet Data frames along with creation date of tweets) and plotted a time series graph, which shows the distribution of sentiments over the period of a year.

  Below plot shows the sentiment of Narendra Modi and Rahul Gandhi for the period of 1 year:

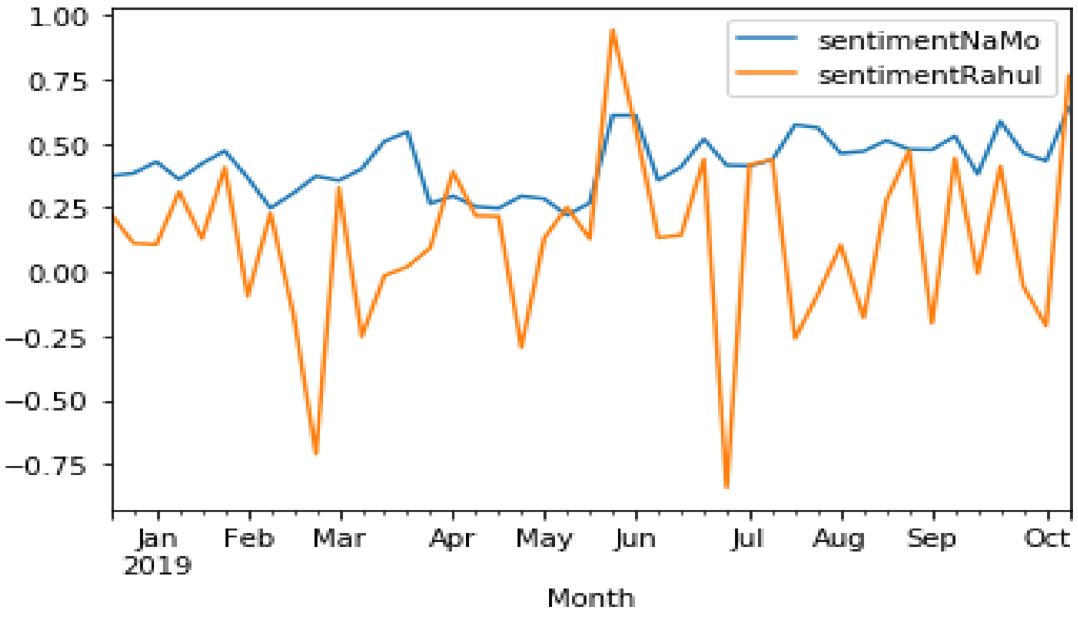


Figure 10- Sentiment of Narendra Modi vs Rahul Gandhi

The above plot shows that, there is a significant difference between the sentiments of both politicians. Sentiments of Narendra Modi is positive and comparatively stationary over the period of time, whereas, there are many peaks (positive and negative) in the sentiments of Rahul Gandhi.

From this, we can infer that, Narendra Modi tweets about the progress and more positive things as a Prime Minister of India. On the contrary, Rahul Gandhi, as a leader of opposition party, beholds tweets with negative emotions referring to issues, challenges or negative events with regards to the leading party.

Event Detection

An analysis of the keywords tweeted by the two political figures was carried to find what makes one less influential compared to another. We have described some keywords from each of their tweets and applied the event detection technique to find how frequently they happen.

What is event detection: The event detection technique requires to know the keywords associated with each event and to assess the minimal count of each word to decide confidently that an event has occurred (source: https://arxiv.org/pdf/1901.00570.pdf).

Assumption: For our analysis, we assume that the event is talked about by the political figures as they happen else the count of keyword associated with that event should be zero.

Limitations: Getting twitter data using the hashtag of the Keywords might better plot the occurrence of the event over a longer time span. However, given the high volume of tweets data, API might fetch data for a very small timeframe or just few days. To avoid this, we used the data fetched from twitter handle of the given person, and it covers the data for nearly one year.

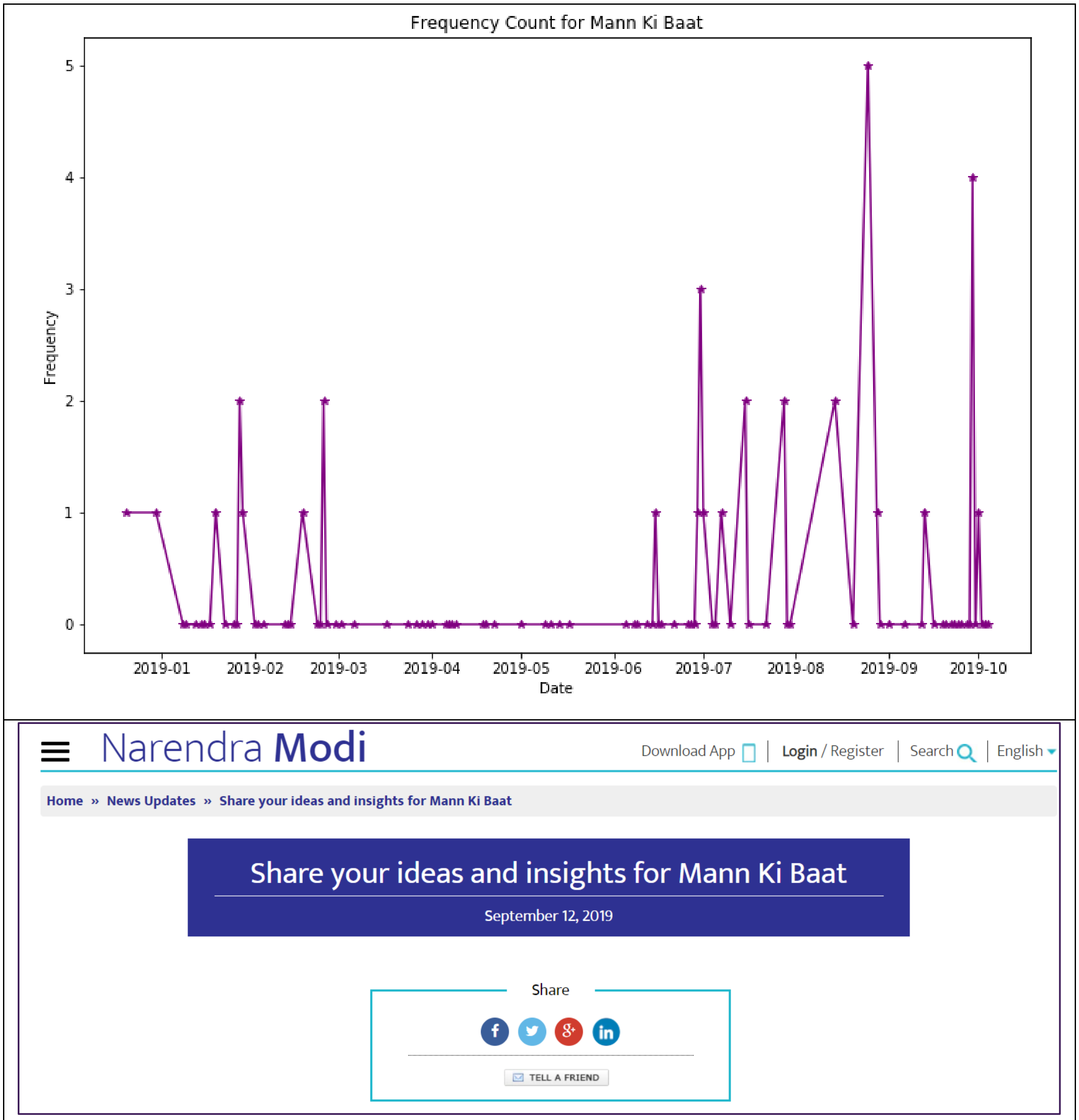
- ### Event detection for keywords used by Narendra Modi:

As we saw earlier that the most frequent words tweeted by Narendra Modi (in the dataset) are “Mann ki Baat”, “Yoga” and “Howdy”.

We further explored the dataset to understand what these words represents and how the keyword “mann ki baat” influence the society at large.

A) **Mann Ki Baat** : According to Wikipedia, Mann Ki Baat (Meaning: "Heart's Voice") is an Indian programme hosted by Prime Minister Narendra Modi in which he addresses the people of the nation on All India Radio, DD National and DD News. In fifteen addresses of Mann ki Baat broadcast so far, more than 61,000 ideas have been received on the website and 1.43 lakh audio recordings by listeners have been received. Each month, some selected calls become a part of the broadcast. He also provided an online platform to share ideas.

From the above Wikipedia definition of ‘Mann Ki Baat’, it was understood to be a monthly occurring event. We tried to find this through the twitter dataset by plotting the time series data for the tweets from Modi containing ‘mann ki baat’. As seen in the chart, Mann ki Baat is almost tweeted every month except for April 2019 and May 2019.



Narendra Modi

Home » News Updates » Share your ideas and insights for Mann Ki Baat

Share your ideas and insights for Mann Ki Baat

September 12, 2019

Figure 11- Event Detection "Mannkibaat"

B) **Yoga Day 2019:**

In June 2019, India celebrated 5th International Day of Yoga 2019 and Prime Minister Narendra Modi performs yoga during a mass yoga event on the 5th International Day of Yoga.

The tweeted dataset was explored to check if “Yoga Day 2019” happens once every year and which month of the year. As seen in the below chart, ‘Yoga’ was most talked about during months June 2019 and July 2019.

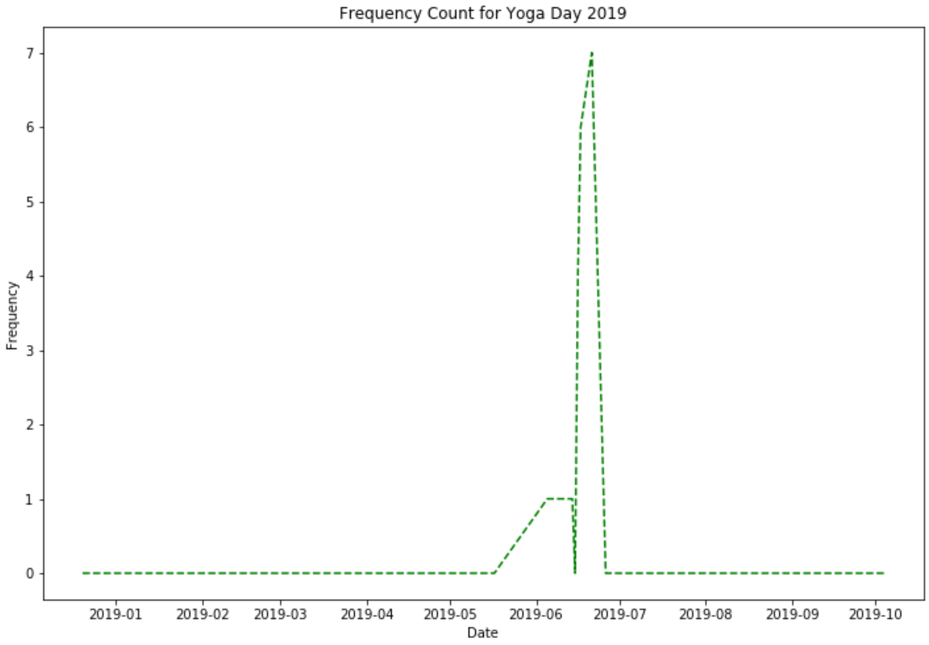
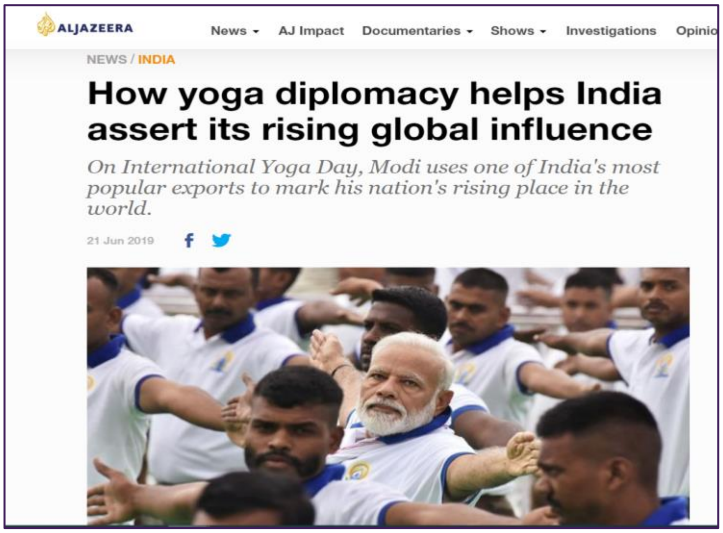


Figure 12- Event Detection "YogaDay2019"



NEWS / INDIA

How yoga diplomacy helps India assert its rising global influence

On International Yoga Day, Modi uses one of India's most popular exports to mark his nation's rising place in the world.

21 Jun 2019

Figure 13- Article mentioning Yoga and global influence

An extract below from a leading Indian Newspaper reads the influence of 'Yoga' internationally.

If China has panda diplomacy, India has yoga, an ancient discipline first practised by Hindu sages thousands of years ago and now one of India's biggest cultural exports.

India's Hindu nationalist Prime Minister Narendra Modi successfully lobbied the United Nations to designate June 21 International Yoga Day in his first year in power in 2014. (Source: https://www.aljazeera.com/news/2019/06/yoga-diplomacy-helps-india-assert-rising-global-influence-190621104210036.html)

C) Howdy

In September 2019, Narendra Modi visited United States of America to address a major rally in Houston called "Howdy, Modi!," where he and US President Donald Trump heaped praise onto one another and touted the strong friendship between their countries.

This showed that Modi has a great international influence/partners in support. We plotted the key word "Howdy" to detect the event and found that it has a burst around end of September 2019. The date can be confirmed from the below picture (source *https://www.businessinsider.com.au/howdy-modi-trump-rally-texas-india-photos-2019-9?r=US&IR=T*)

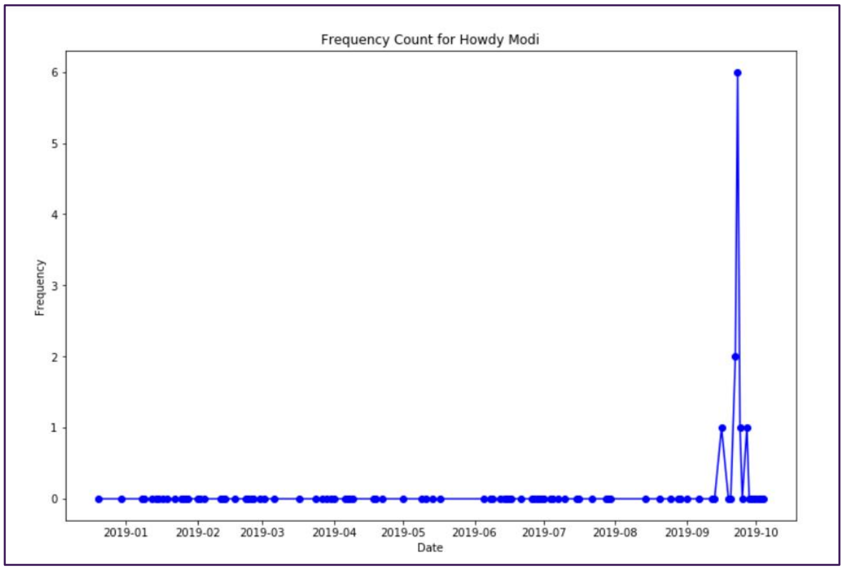


Figure 14- Event Detection "HowdyModi"

REUTERS/Daniel Kramer

U.S. President Donald Trump and Indian Prime Minister Narendra Modi during a 'Howdy, Modi' rally celebrating Modi at NRG Stadium in Houston, Texas, U.S. September 22, 2019.

Figure 15- HowdyModi Event in Huston

- ## Event detection for keywords used by Rahul Gandhi:

We will analyze the key words tweeted by Rahul Gandhi to show that he is less influential as he talks less about driving economy & democracy and criticizes the opposition more. This reflects less positive on his personality.

The most tweets keyword by Rahul Gandhi is "Rafael". The 'Rafale' deal controversy is a political controversy in India related to the purchase of 36 multirole fighter aircraft for a price estimated at €7.8 billion by the Defense Ministry of India from France's Dassault Aviation.

Rahul has been very vocal about it holding it as incorrect decision. The event was detected around November 2018 as seen in the twitter plot below. This can be confirmed from the picture below(from source https://www.news18.com/news/politics/dassault-ceo-is-lying-to-protect-pm-modi-in-rafale-scam-says-rahul-gandhi-1927619.html).

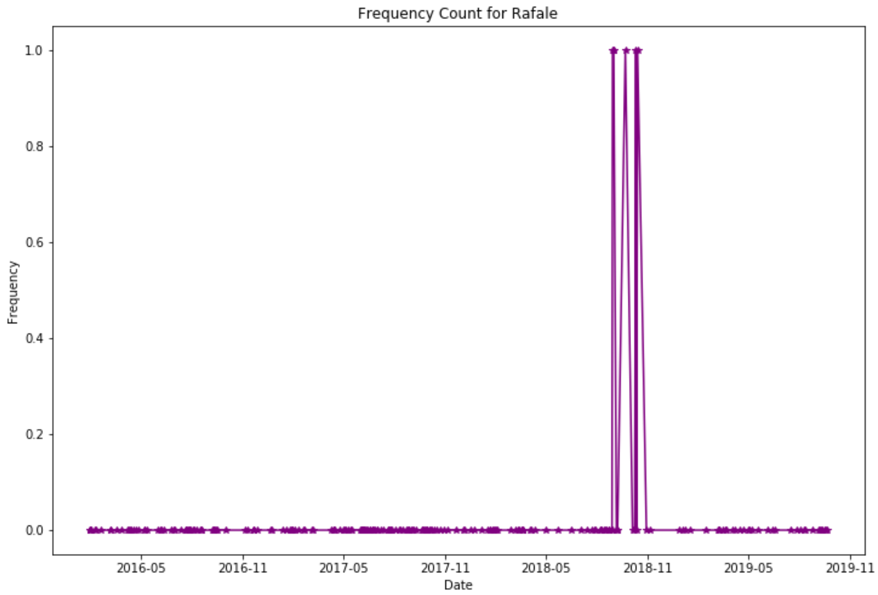
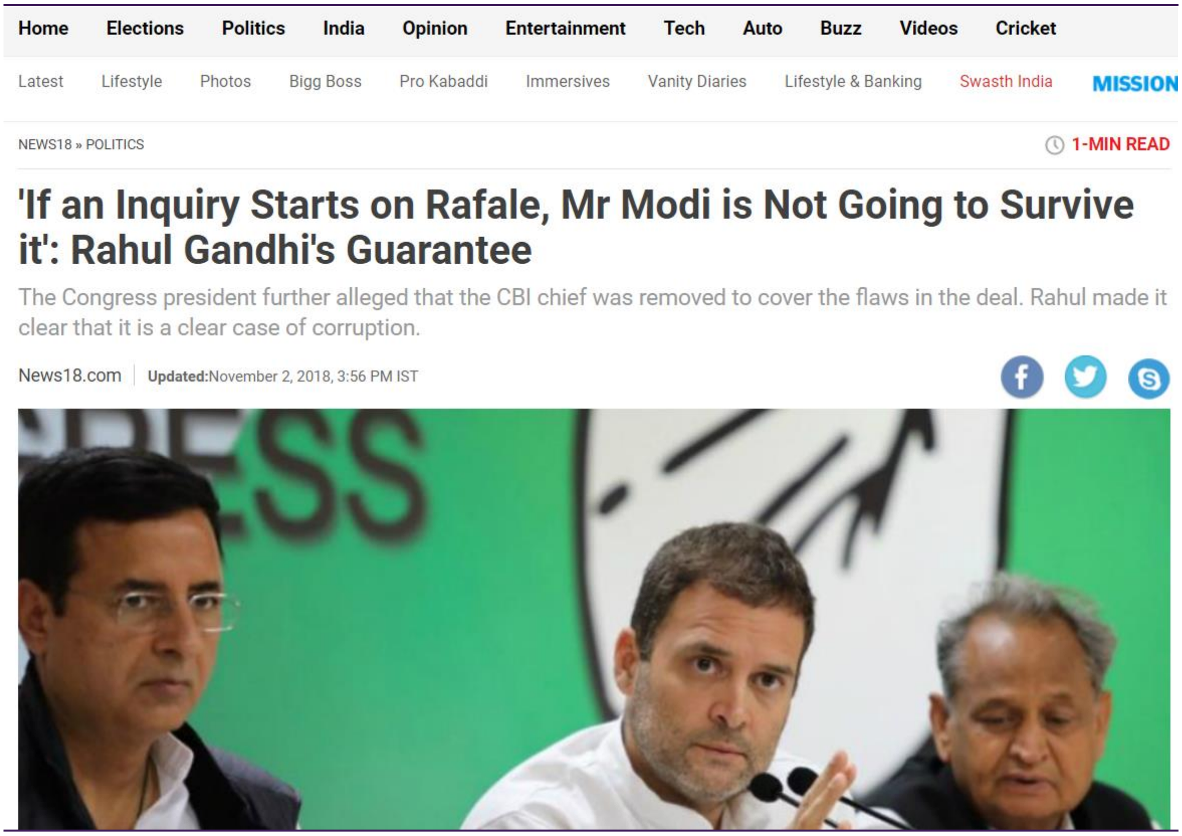


Figure 16- Event detection "rafale"



NEWS18 » POLITICS

1-MIN READ

'If an Inquiry Starts on Rafale, Mr Modi is Not Going to Survive it': Rahul Gandhi's Guarantee

The Congress president further alleged that the CBI chief was removed to cover the flaws in the deal. Rahul made it clear that it is a clear case of corruption.

News18.com | **Updated:**November 2, 2018, 3:56 PM IST

Figure 17- Rafale news

Social Network Analysis (SNA)

- ## Network Visualization using graph:

Data Collection and limitation: Both the political figures have followers in millions. Narendra Modi has ~ 50 million whereas Rahul Gandhi has nearly ~11 million followers. It was not possible using twitter API free developer account to obtain the data. We have sampled the data in a certain ratio of their followers and followed to represent the population data.

The friends/followers of followers data couldn't be extracted as twitter API was throwing error as "Not Authorized" suggesting private accounts.

Also, it was not possible to clearly plot too many points in the network graph, we have taken minimum nodes which can be seen clearly in the graph using the open source tool "Gephi".

Purpose of SNA analysis:

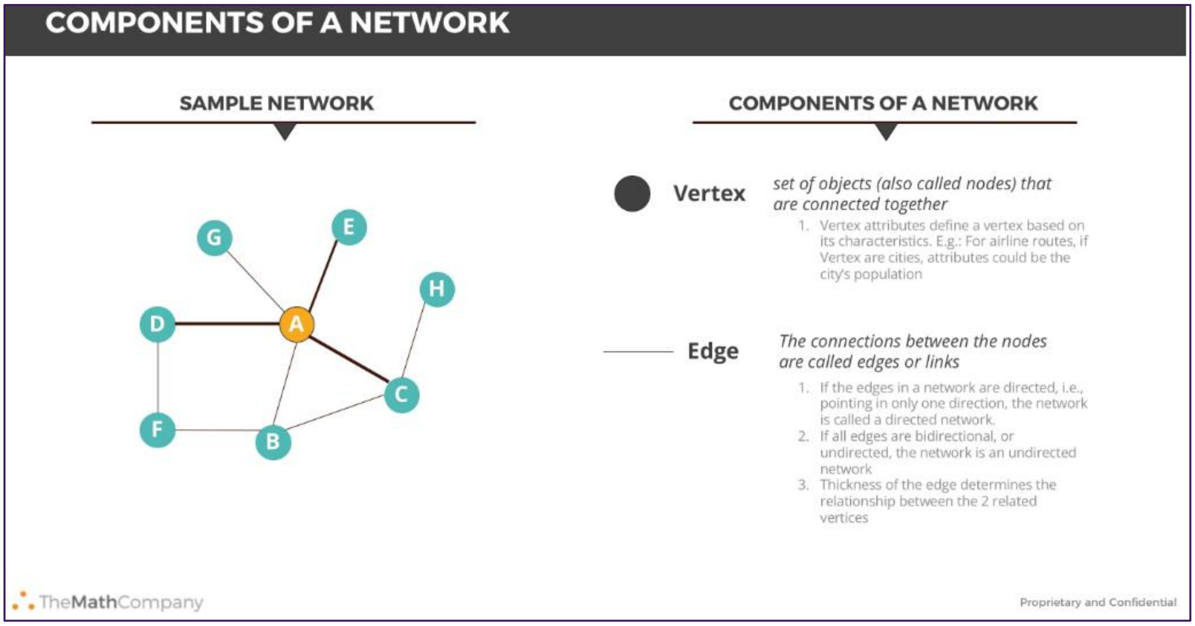
Networks are everywhere, networks of roads, a network of friends and followers on social media, and a network of office colleagues. They play a significant role from spreading useful information to influencing national elections. Through the SNA analysis we will show that the volume of followers of Narendra Modi are much higher compared to the volume of followers of Rahul Gandhi. Also, we used "Centrality Measures" to find "quantitatively how much" one person is more central/important compared to another.

Methodology for SNA measures:

Tools Used: Python Networx library and open source tool Gephi.

We used directed graph plotted in 'Gephi' to show the relationships of followed and followers with the given person.

A graph constitutes edges and nodes as shown in the picture below. A directed graph has arrow at the head to represent who is following whom.



COMPONENTS OF A NETWORK

SAMPLE NETWORK

COMPONENTS OF A NETWORK

Vertex *set of objects (also called nodes) that are connected together*

1. Vertex attributes define a vertex based on its characteristics. E.g.: For airline routes, if Vertex are cities, attributes could be the city's population

Edge *The connections between the nodes are called edges or links*

1. If the edges in a network are directed, i.e., pointing in only one direction, the network is called a directed network.
2. If all edges are bidirectional, or undirected, the network is an undirected network
3. Thickness of the edge determines the relationship between the 2 related vertices

Figure 18- Network Components

In the below social network graph using twitter data, the black line directed towards the center (Narendra Modi) represents the followers. And, the red lines directed away from the center represents the followed.

As visible for the 2 graphs, graph 1 has a greater number of black lines over red compared to graph2.

This can be interpreted as Narendra Modi having more followers compared to Rahul Gandhi.

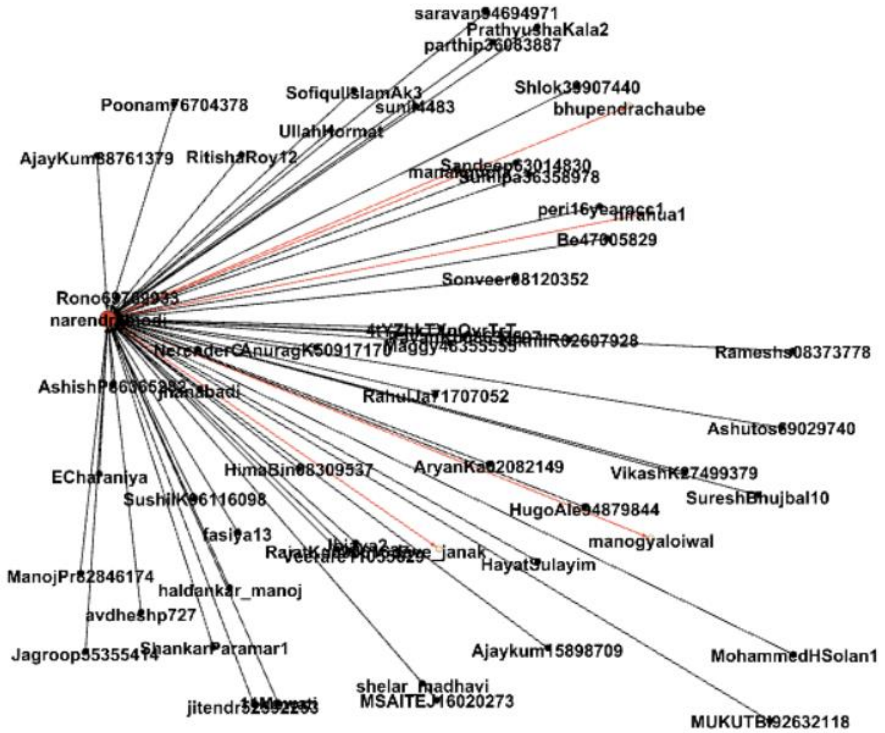


Figure 19- Directed Social Network Graph for Narendra Modi

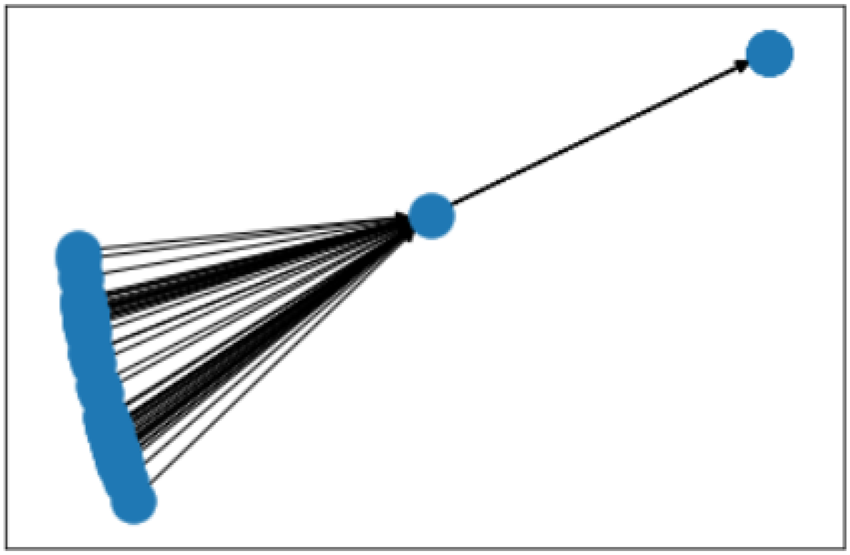


Figure 20- The same network Graph 1 above was plotted using Python NetworkX

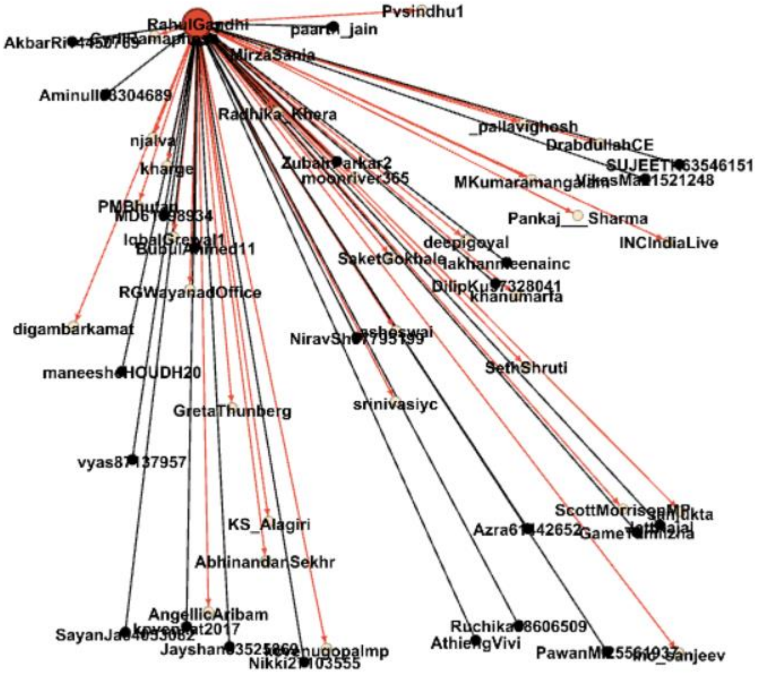


Figure 21- Graph 2: Directed Social Network Graph for Rahul Gandhi

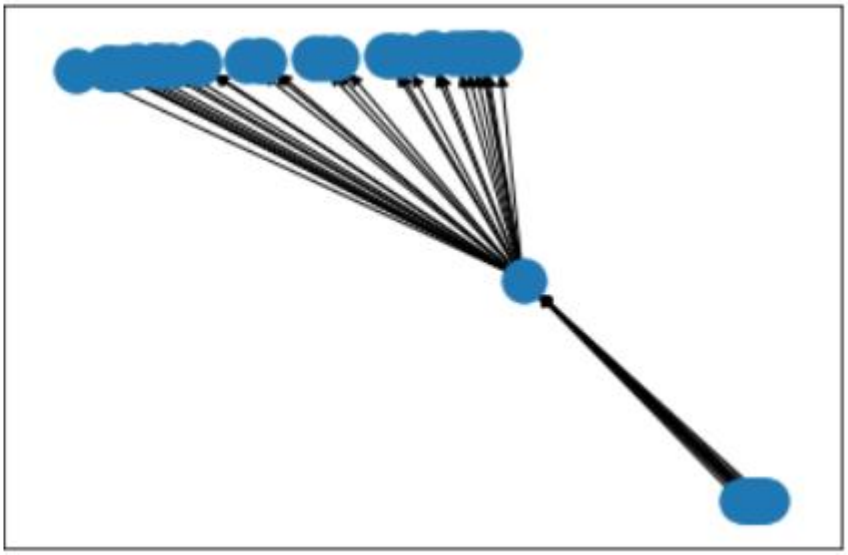


Figure 22- The same network Graph 2 above was plotted using Python NetworkX

- ## Comparing Centrality: Katz Centrality Measure

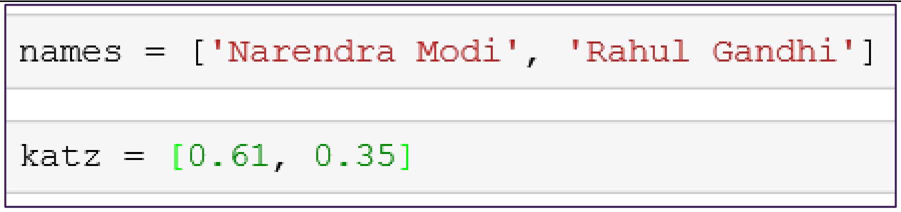
In graph theory, the Katz centrality of a node is a measure of centrality in a network. It was introduced by Leo Katz in 1953 and is used to measure the relative degree of influence of a node within a social network.

Reason to choose Katz Centrality over other Centrality measures:

Katz centrality computes the relative influence of a node within a network by measuring the number of the immediate neighbors (first degree nodes). Connections made with distant neighbors are, however, penalized by an attenuation factor alpha.

Since, we have only first-degree nodes (followed and followers) due to twitter API limitations, we found Katz Centrality as the best fit.

The below bar chart compares the Katz Centrality for Narendra Modi and Rahul Gandhi. The Katz centrality for Narendra Modi is 0.61 and for Rahul Gandhi is 0.35.



```
names = ['Narendra Modi', 'Rahul Gandhi']

katz = [0.61, 0.35]
```

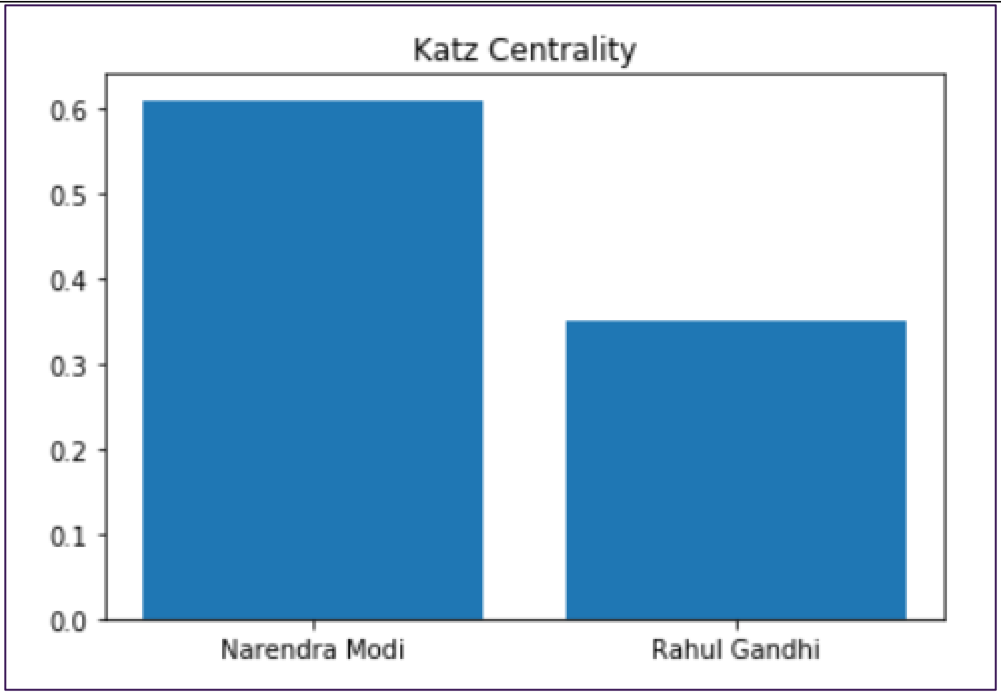


Figure 23- Katz Centrality

Conclusion

We mentioned in the introduction part that social media has become widely used platform for politician to communicate with wider population. For this assignment we took data from Twitter but there are other sources such as Facebook, Reddit, Web Forums, Blogs, etc., from where data can be extracted and analyzed.

From the data exploration, sentiment analysis, event detection and social network measure analysis we conclude that Narendra Modi stands out in terms of his influence over the people of largest democracy in the world. He exudes positive vibes/sentiments when he tweets. He encourages the youth by inviting their ideas to innovate and boost the Indian economy. He is friendly with other political leaders across the globe. In his social network he has large followers base and he has high centrality compared to Rahul Gandhi.

The above results could be analysed further with the data from other data sources and calculating different SNA and Centrality measures.

References

- Rmit.instructure.com. (2019). MyApps Portal. [online] Available at: https://rmit.instructure.com/courses/49836/pages/week-2-learning-materials-slash-activities?module_item_id=1802884 [Accessed 1 Sep. 2019].
- Stieglitz, Stefan & Dang-Xuan, Linh. (2014). Social Media and Political Communication - A Social Media Analytics Framework. Social Network Analysis and Mining. 3. 1277-1291. 10.1007/s13278-012-0079-3.