CSE 472: Social Media Mining

PROJECT 1: Social Media Data Analysis

by Sahil Vora

PROJECT OBJECTIVE:

To crawl Twitter data on the specific topic of COVID-19 Vaccine as well as process and perform exploratory analysis on extracted data.

PROJECT OVERVIEW:

Brief Summary of the Project

Based on the given choices for the project, here is a summary of steps done to get the results which are explained section wise:

1. The **API method** was selected using the **Tweepy** Library.
2. The data that was collected was regarding the two attitudes towards COVID-19: **anti-vaccine and pro-vaccine.**
3. The networks created using this crawled data are **Diffusion Network** and **Word Co-occurrence** networks created using the library **networkx** and **natural language toolkit (nltk)** is used for the later.
4. To show a contrast between the networks for anti-vaccine and pro-vaccine the same networks are also created for a **larger dataset** collected.

How was the Data Scraped

Twitter data can be scraped with the help of APIs provided by creating a developer account and requesting for an elevated access. The twitter API that are used for this project are version 2 APIs and post requesting the same using an **API method** and with the help of the library **Tweepy**, the data was fetched based on the queries having the hashtags as the search term. Here is step by step explanation on how the data was scraped.

1. Tweepy provides cursor methods that helps in fetching data in bulk and using this function the API calls the ***search_tweets*** method to search for the hashtags. The number of tweets that are fetched for the smaller dataset for **each hashtag** is approximately **30 and 100** for the larger dataset.

2. Since this query method also requires a date after which the tweets should be fetched so the date of **1st April 2021** was chosen as this was the time around which vaccines were introduced around the world.
3. Since with elevated access from twitter developer portal, we can only fetch **180 requests in 15 minutes** with default response of 10 to max 100 results per response, the crawler script was made to sleep for 15 minutes after max limit was encountered and was resumed thereafter.
4. The search terms for Anti-vaccine campaign such as ***#StopVaccination*** were picked from the repository mentioned. Similarly Pro-Vaccine campaign's hashtags were also picked in the same method. On dry run for all the hashtags, some of the hashtags don't result in the desired number of tweets and hence were dropped. These hashtags are stored in ***AntiVaccineKeywords.txt*** and ***ProVaccineKeywords.txt*** and can be updated anytime. Here is a snapshot on how the keywords look like for both the campaigns.

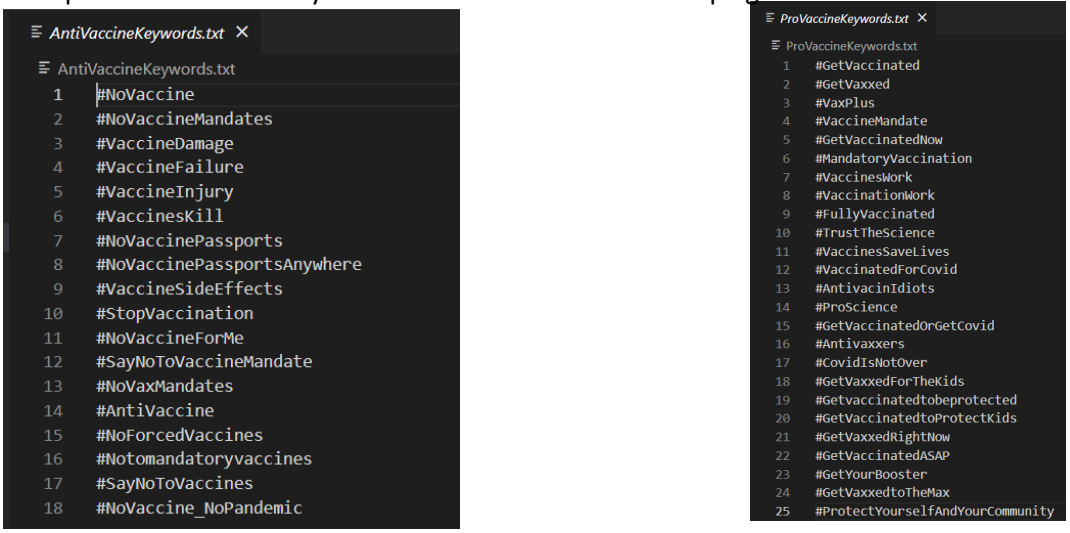

Figure 1. Anti-Vaccine Search hashtags, Pro-Vaccine Search Hashtags (left to right)

5. The data was stored in the JSON format. The ***search_tweets*** query results in data object of Tweets which contains a JSON object that has been stored in ***AntiVaccineTweets.json*** and ***ProVaccineTweets.json*** and the larger dataset is stored in ***AntiVaccineTweets1.json*** and ***ProVaccineTweets1.json***
6. This entire process is done in the script of **Project_Scraping_Script.py**

Data Cleaning and Collection

1. For Diffusion Network

1.1. After the data was collected, for diffusion network the data needs to be refined and presented in the form of pandas data frame which is easier to use with networkx.

1.2. Since the JSON object had many attributes, the ones that are useful for the network creation are:

created_at, id, in_reply_to_screen_name, in_reply_to_status_id, in_reply_to_user_id,retweeted_id, retweeted_screen_name, user_mentions_screen_name, user_mentions_id, full_text, user_id, screen_name, followers_count

1.3. In the diffusion network each node is a user and if there is any user mention, reply for retweet forming the **Tweet Propagation Network.**
1.4. For this the data is cleaned using the mentioned attributes and a new data frame named *tweets_new_df* is used.
1.5. This is filled with the information of the user itself using screen name and id.
1.6. Then user mentions are filled with the help of ***user_mentions*** attributes.
1.7. Similarly, data is fetched for replies and the retweets using the attributes mentioned.
1.8. The **interactions** i.e., the edges are formed if there is any kind of relation between each of the tweets.
1.9. And finally, the data was cleaned by removing the rows having no interactions.

2. For Word Occurrence Network

2.1. For word Occurrence, the text of the tweets was needed to be cleaned.
2.2. All the URLs in the texts were removed.
2.3. The entire text was converted to lowercase.
2.4. Also, the stop words such as (a, an, the, etc.) were removed with the help of **nltk** library.
2.5. Since we need words that occur together, we utilised the ***bigrams*** feature of nltk library.
2.6. Using these bigrams, a network of word-occurrence network was created.

Social Media Network Construction and Visualisation

1. Diffusion Network

*These following network and graphs are created using the **larger dataset**.*

1.1. Anti-Vaccine Attitude

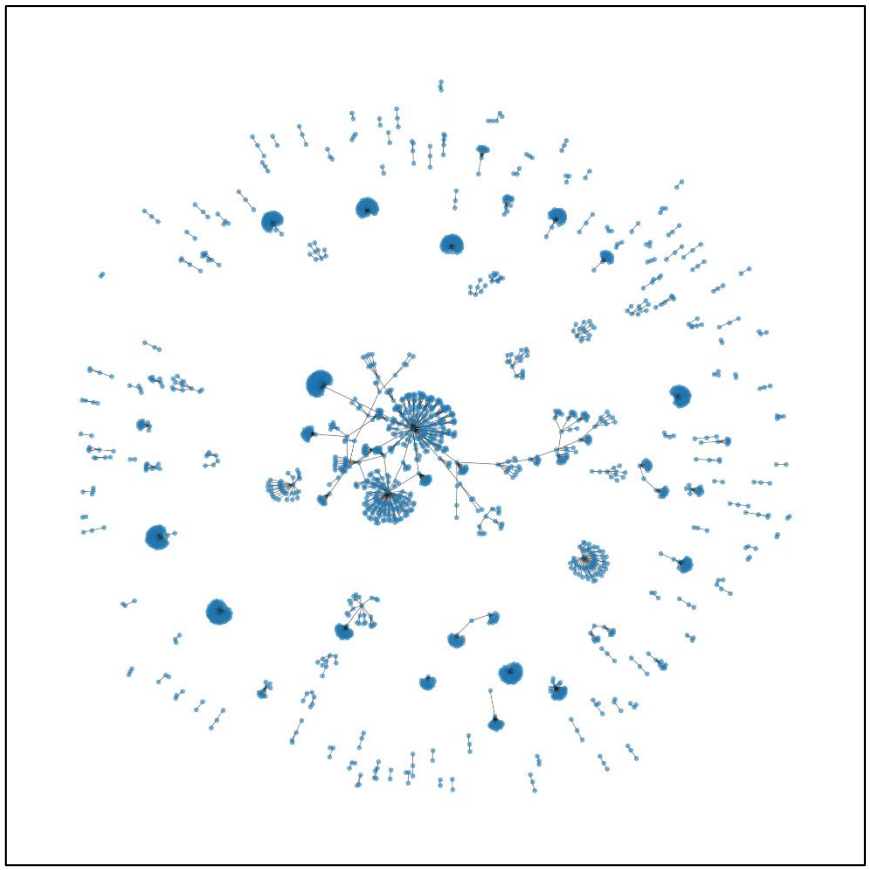

Figure 2. Diffusion Network created for Anti-Vaccine Attitude

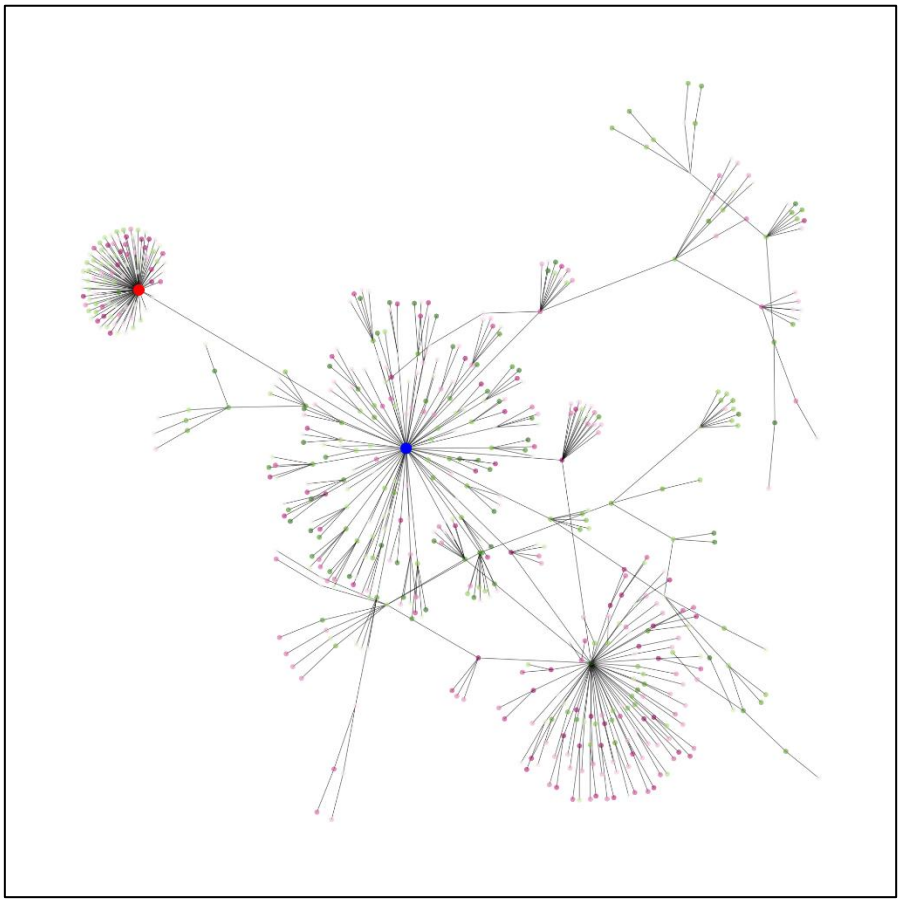

Figure 3. Largest Connected Subgraph for Anti-Vaccine Attitude

1.2. Pro-Vaccine Attitudes

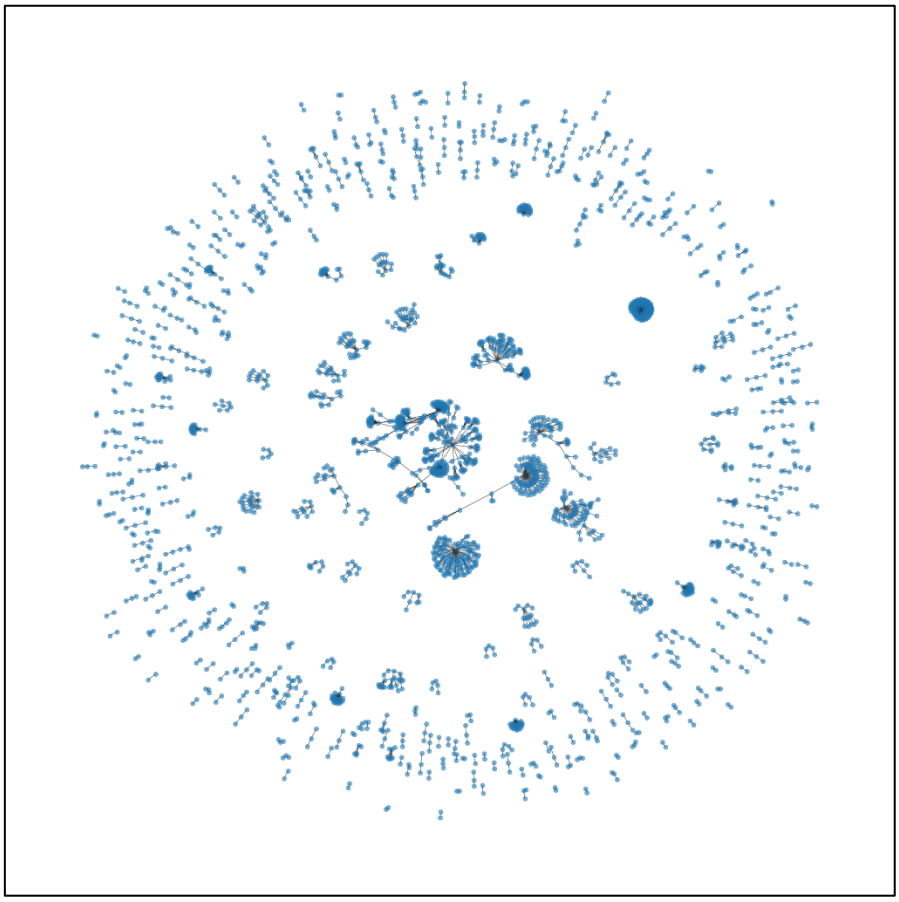

Figure 4. Diffusion Network for Pro-Vaccine Attitude

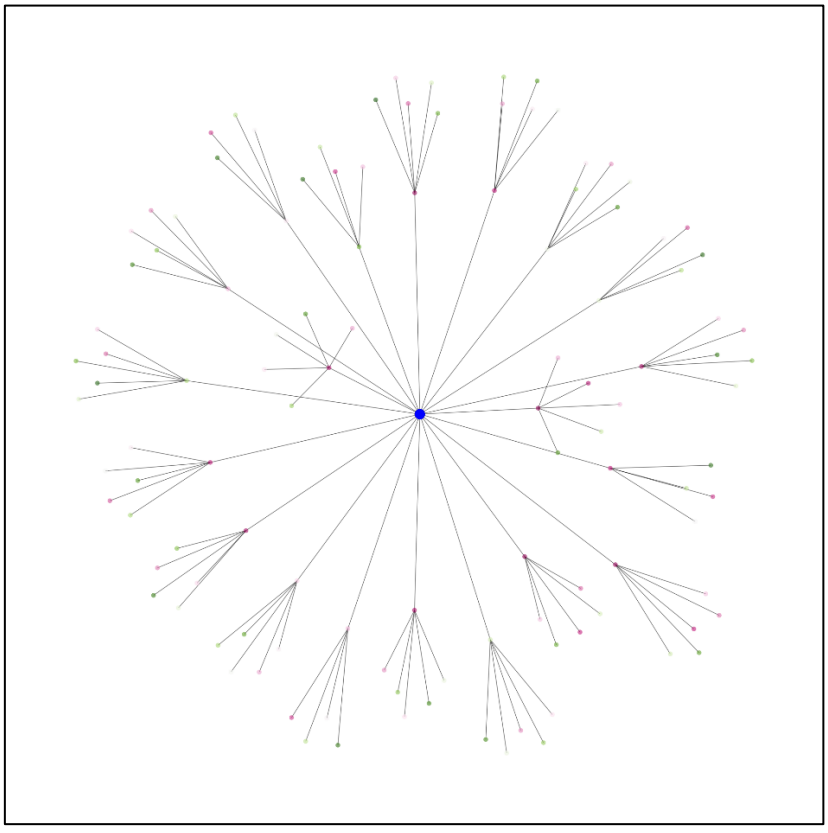

Figure 5. Largest Connected Subgraph for Pro-Vaccine Attitude

2. Word Occurrence Network

2.1. Anti-Vaccine

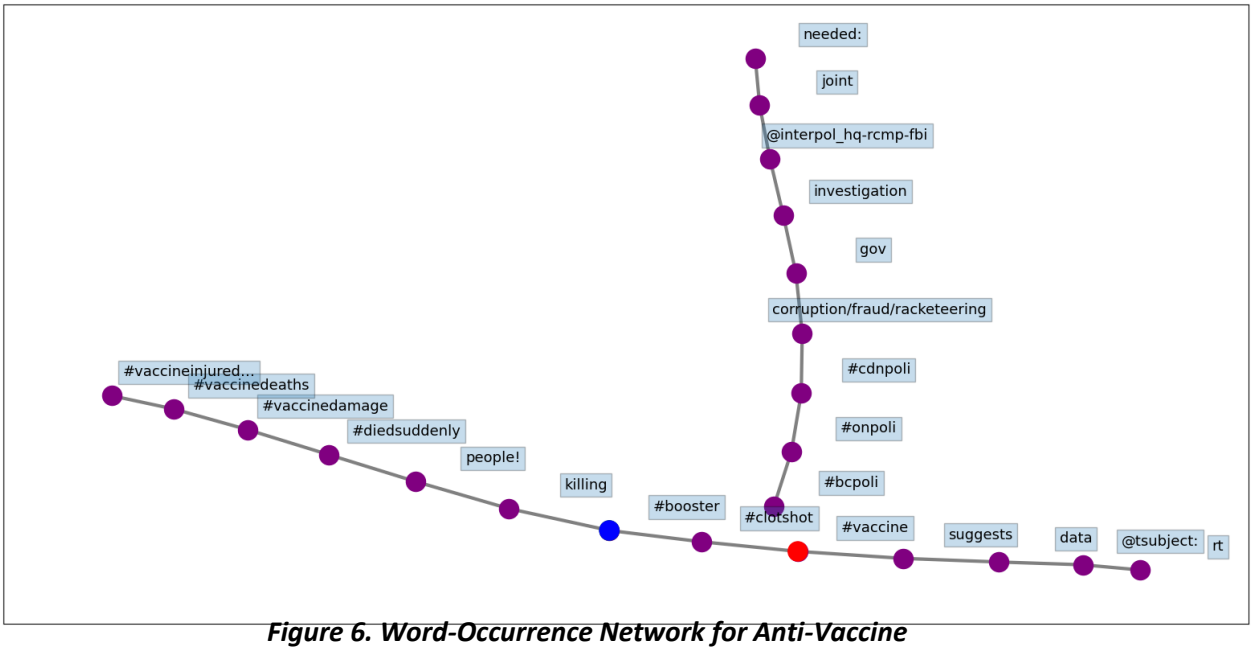

Figure 6. Word-Occurrence Network for Anti-Vaccine

2.2. Pro-Vaccine

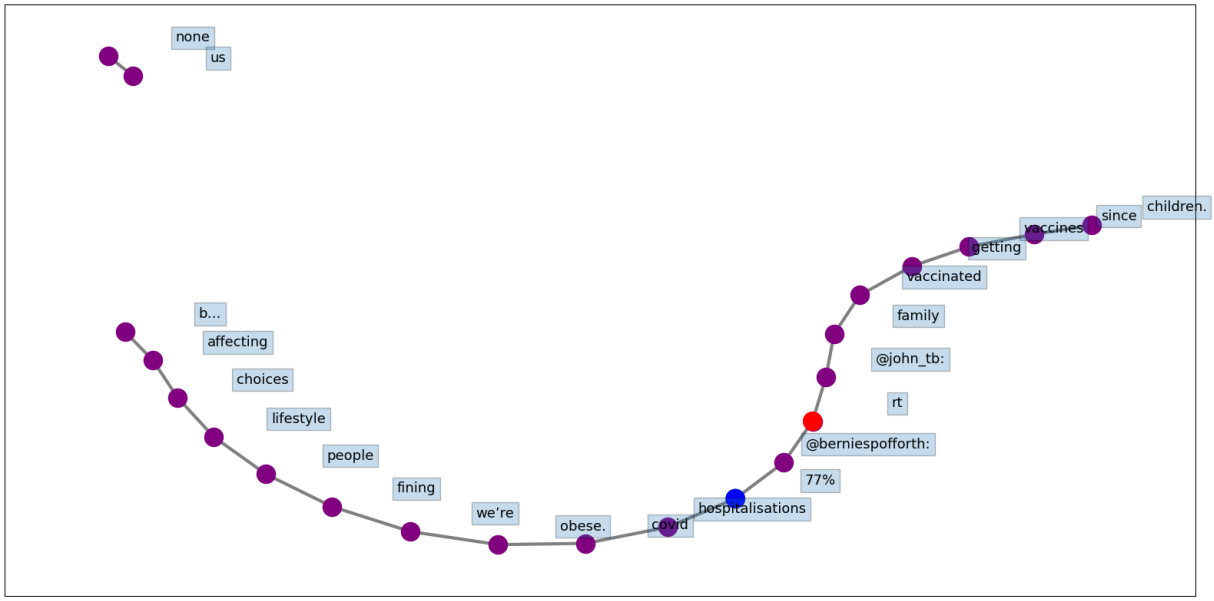

Figure 7. Word-Occurrence Network for Pro-Vaccine

Network Measure Calculation Results

	DIFFUSION NETWORK GRAPH	
NETWORK MEASURE	**ANTI-VACCINE**	**PRO-VACCINE**
Number of Nodes	1994	2465
Number of Edges	1846	2034
Maximum Degree of Graph	93	90
Minimum Degree of Graph	1	1
Average Degree of Graph	1.9	1.7
Most Frequent Degree of nodes	1	1
Connected Graph?	No	No
Number of connected components	156	447
Maximum Degree Centrality	0.05	0.04
Maximum Closeness Centrality	0.11	0.04
Maximum Betweenness Centrality	0.07	0.00
Maximum Page Rank	0.02	0.02
Diameter of largest connected graph	14	4

	LARGEST CONNECTED SUBGRAPH	
NETWORK MEASURE	**ANTI-VACCINE**	**PRO-VACCINE**
Number of Nodes	573	116
Number of Edges	577	115
Connected Graph?	Yes	Yes
Diameter	14	4
Average Distance between any two nodes	4.84	3.54
Maximum Degree Centrality	0.16	0.17
Maximum Closeness Centrality	0.38	0.55
Maximum Betweenness Centrality	0.86	0.96
Maximum Page Rank	0.07	0.07

	WORD OCCURRENCE NETWORK	
NETWORK MEASURE	**ANTI-VACCINE**	**PRO-VACCINE**
Maximum Degree Centrality	0.10	0.10
Maximum Closeness Centrality	0.16	0.17
Maximum Betweenness Centrality	0.17	0.43
Maximum Page Rank	0.05	0.05

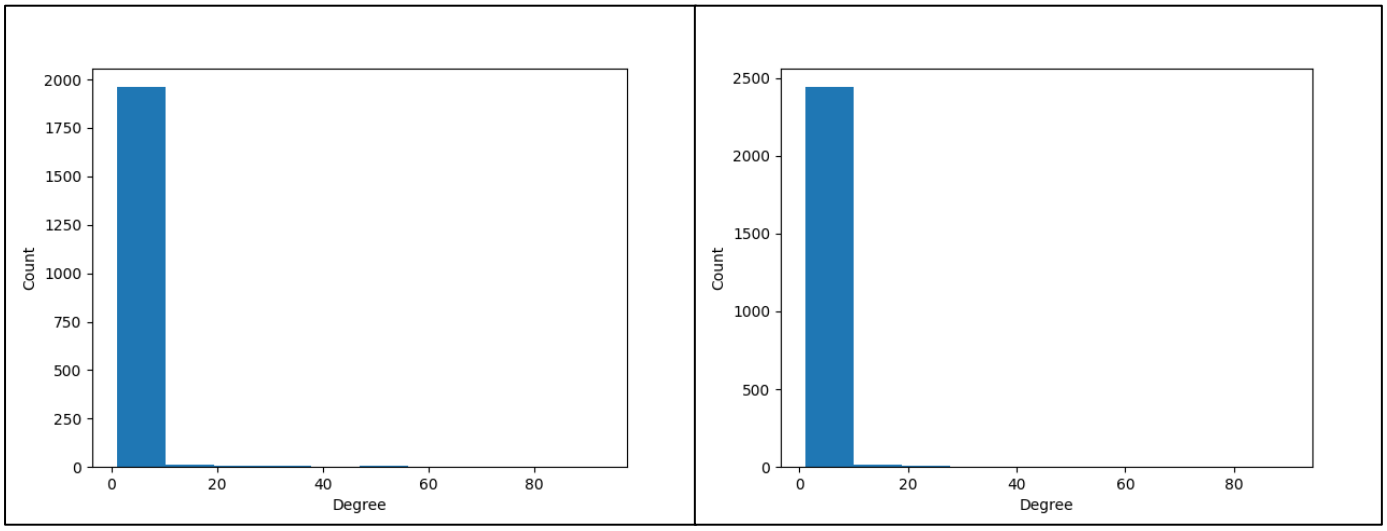

Figure 8. Degree Distribution for diffusion network of Anti-Vaccine(left) and Pro-Vaccine(right)

Social Media Network and Measures for smaller datapoints

*These following network and graphs are created using the **smaller dataset**.*

1. ANTI-VACCINE

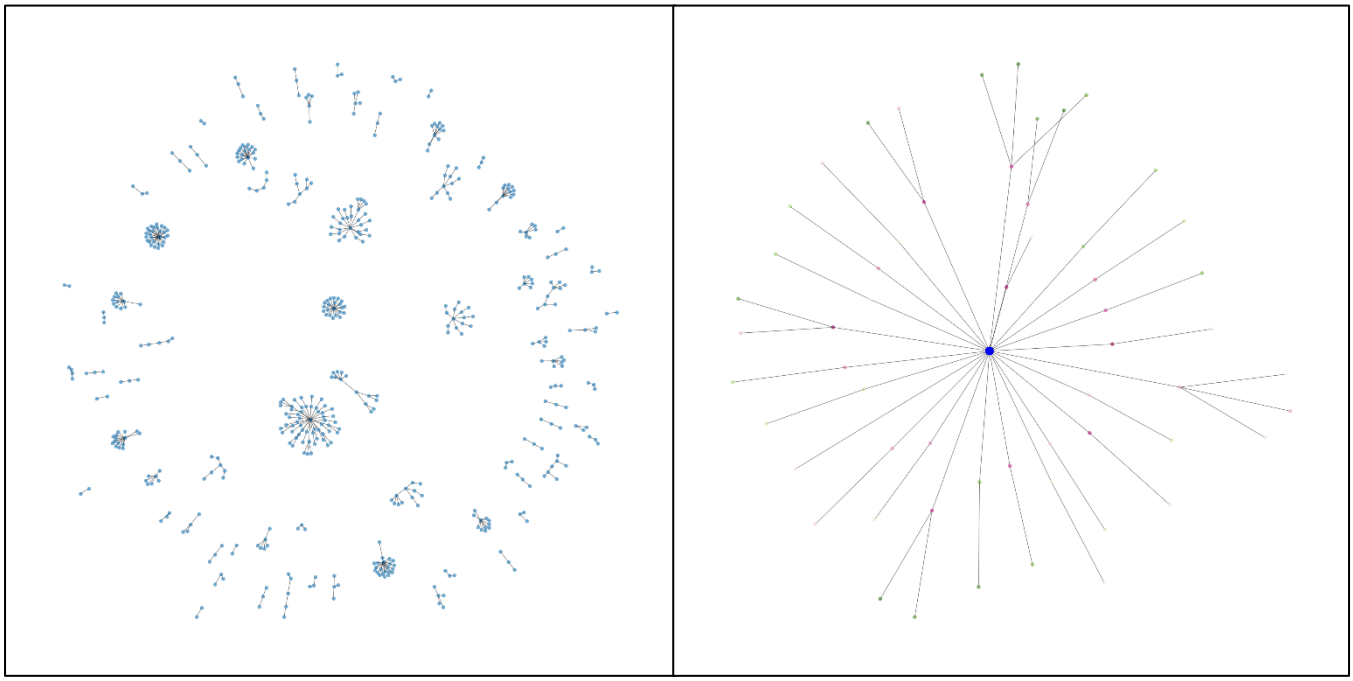

Figure 9. Diffusion network(left) and Largest Connected Sub-Graph for Anti-Vaccine

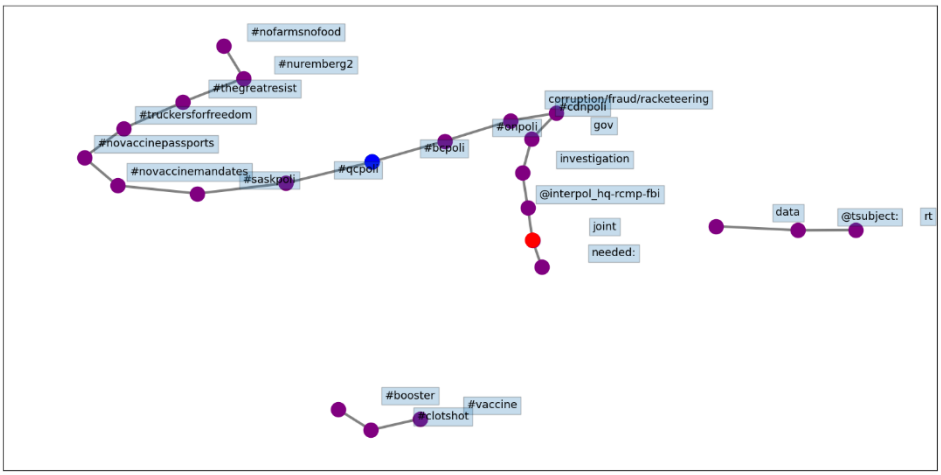

Figure 10. Word-Occurrence Network for Anti-Vaccine

2. PRO-VACCINE

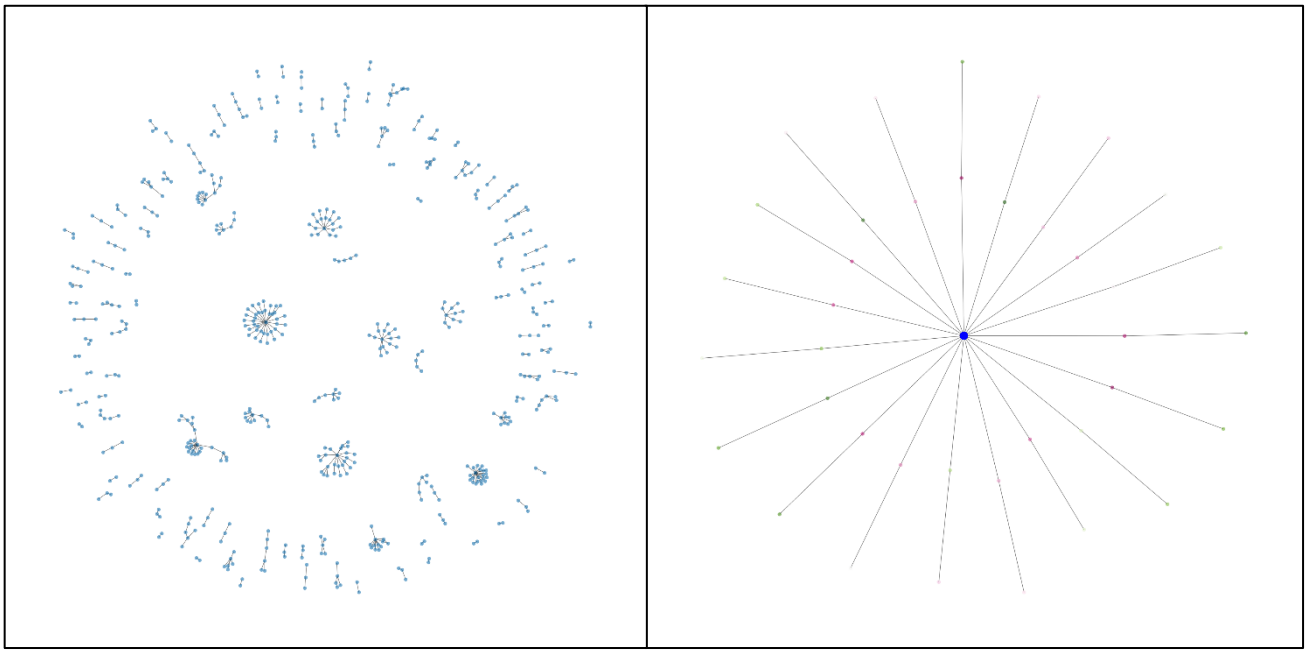

Figure 11. Diffusion network(left) and Largest Connected Sub-Graph for Pro-Vaccine

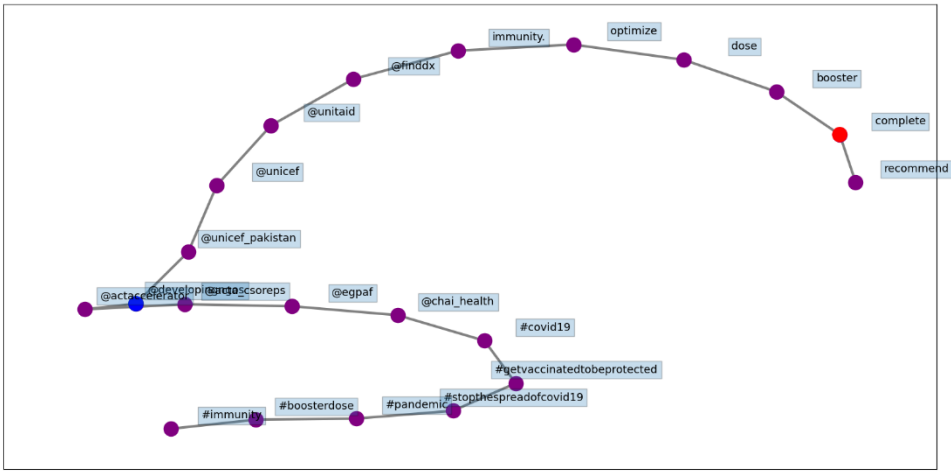

Figure 12. Word-Occurrence Network for Pro-Vaccine

	DIFFUSION NETWORK GRAPH	
NETWORK MEASURE	**ANTI-VACCINE**	**PRO-VACCINE**
Number of Nodes	526	570
Number of Edges	449	445
Maximum Degree of Graph	27	21
Minimum Degree of Graph	1	1
Average Degree of Graph	1.7	1.6
Most Frequent Degree of nodes	1	1
Connected Graph?	No	No
Number of connected components	77	128
Maximum Degree Centrality	0.05	0.04
Maximum Closeness Centrality	0.07	0.04
Maximum Betweenness Centrality	0.01	0.00
Maximum Page Rank	0.02	0.02
Diameter of largest connected graph	4	4

	LARGEST CONNECTED SUBGRAPH	
NETWORK MEASURE	**ANTI-VACCINE**	**PRO-VACCINE**
Number of Nodes	59	39
Number of Edges	58	38
Connected Graph?	Yes	Yes
Diameter	4	4
Average Distance between any two nodes	3.03	2.87
Maximum Degree Centrality	0.43	0.50
Maximum Closeness Centrality	0.64	0.67
Maximum Betweenness Centrality	0.97	0.97
Maximum Page Rank	0.18	0.22

	WORD OCCURRENCE NETWORK	
NETWORK MEASURE	**ANTI-VACCINE**	**PRO-VACCINE**
Maximum Degree Centrality	0.09	0.10
Maximum Closeness Centrality	0.16	0.18
Maximum Betweenness Centrality	0.28	0.53
Maximum Page Rank	0.06	0.06

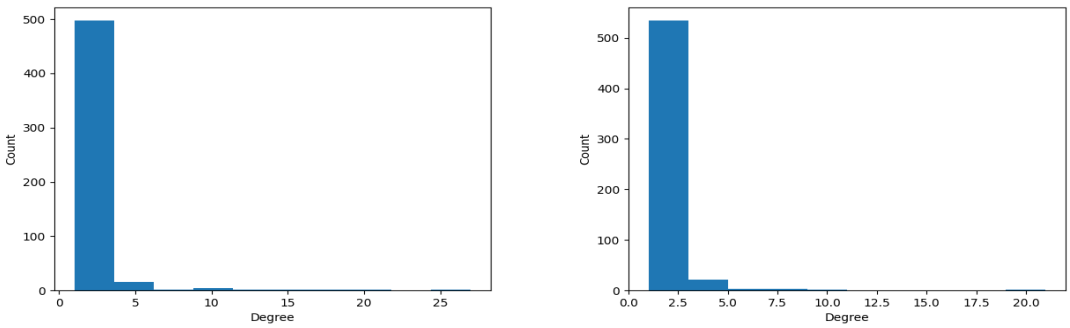

Figure 13. Degree Distribution for diffusion network of Anti-Vaccine(left) and Pro-Vaccine(right)

Observation and Characteristic Differences

Based on the graphs and network formed, it is very evident that in the diffusion network, the tweets are connected densely with some tweets clearly originating as the centre of the network and other tweets emerging from these centre points for the Anti-Vaccine campaign. There are very small, connected components which are also present. While for the Pro-Vaccine campaign and tweets, the centre or origin point is singular and not very densely populated. Here is a side-by-side comparison of the largest connected graph of the diffusion network which clearly depicts these observations.

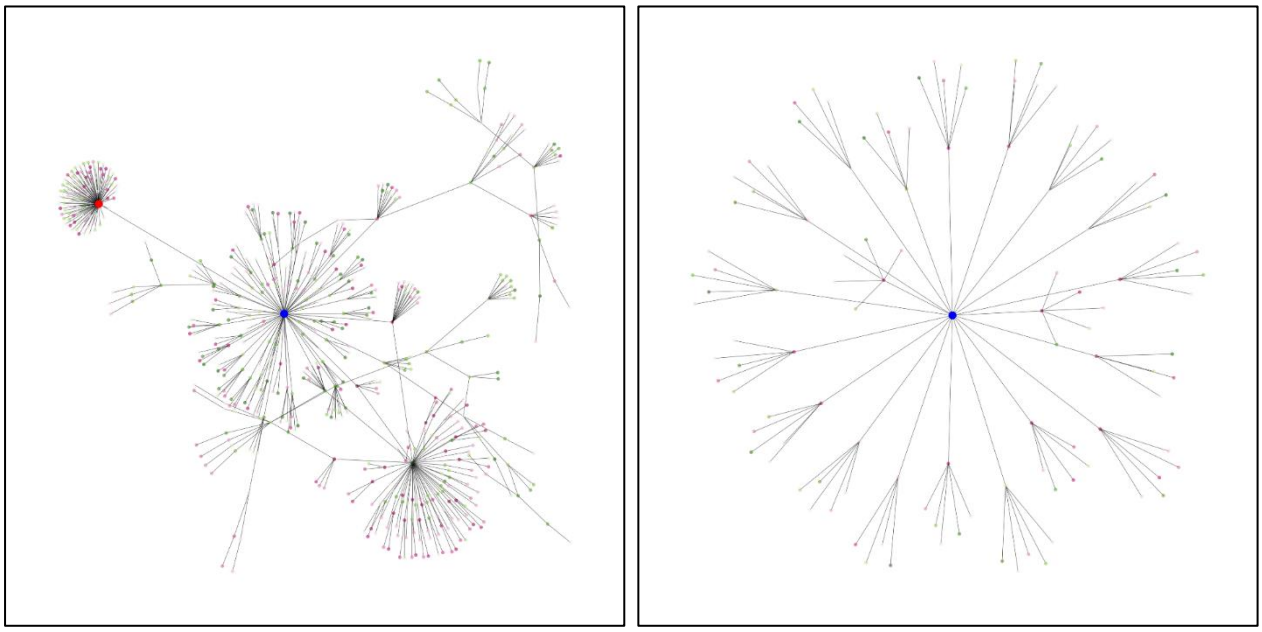

Figure 14. Largest Connected diffusion network of Anti-Vaccine(left) and Pro-Vaccine(right)

Summary

The project helps in clear understanding on how to create social networks along with how data available on twitter can resemble closed connections and proves evidence of connections and networking for campaigns that promotes a common motive. Anti-Vaccine campaign of tweets shows stronger opinion and is densely connected from the data obtained.

Reference

1. How to download and visualise your Twitter Network – Steve Hedden – Towards Data Science.
2. Analysis of Twitter Social Network – Pratik Parija - Social Media: Theories, Ethics, and Analytics.
3. Analyse Co-Occurrence and Networks of words Using Twitter Data and Tweepy in Python – Earth Data Analytics – Earth Lab