NETWORK ANALYSIS & INSIGHTS OF COVID-19 TWEETS

Social Information and Network Analysis
Autumn 2021 | University of Technology Sydney

Enriquez, Joe Drigo
Saadoun, Mohammad
Sampe, Irfan
Tin, Hnin Pwint
Wong, Ki Ming

Network Analysis & Insights on COVID-19 Tweets

Abstract (Executive Summary)

The COVID-19 pandemic swept the world globally at the early parts of 2020 and continues until this day. Twitter, often touted as a modern indicator of the general public's sentiment, was a central platform in which conversations and discourse about the virus are taking place. We endeavored to analyse these tweet-based discussions with the use of network analysis techniques. These have provided the following main insights on 3 specific areas that we have explored: (1) When **communities** form on Twitter, the phenomenon of social media acting as an 'echo chamber' becomes more apparent – where we found Twitter users interacting mainly with people who already share their opinion. (2) the most **central** accounts are not always the names we expect. Although we see public officials (e.g. therealdonaldtrump) and news channels (e.g. CNN) in the list of central accounts, we also found that 'regular' people with less followers (e.g. jcho710 with 10,000 followers) can lead the discussion with enough activity and tweet engagement. (3) the **sentiment** on COVID-19 is more nuanced than first glance. Although 'fear' and 'sadness' still account for a significant number of tweets, they are not the overwhelming majority in emotions. Tweets that are 'happy' and 'surprised' are also seen in the discussions. With these findings, the team hopes to shed clarity on preconceived notions we may have concerning this incredibly complex period in the world.

Introduction: Research Questions and Data Sources

Our Research Questions

In 2020, the world was gripped with the spread of the COVID-19 virus, imploring the World Health Organisation (WHO) to declare a state of pandemic globally (Cucinotta & Vanelli, 2020). Throughout last year, a key part of the conversation is the arrival and readiness of the vaccines – which experts identify is the prime method by which we can control the virus permanently (Karron, 2020).

Unlike previous pandemics, today's society permeated by the functionalities of the internet have several ways to converse about the pandemic and the readiness of the vaccines eagerly awaited by many. Twitter is an important platform by which these conversations happen. The nature of tweeting, which are essentially real-time, short 280-character broadcasts, has become an important source of information and are generally considered a great indicator of public sentiment and opinion (Wicke & Bolognesi, 2021).

With this, we hope to take a look at datasets containing tweets related to the COVID-19 virus and vaccine deployment from 2020, with intent to see how networks form around this incredibly critical conversation, and see if there are central actors, communities formed, and gain a general overview of the sentiment of the masses with regard to the virus and the vaccine over time.

Our Research Question: What are valuable insights we can get from analysing the network of COVID-19-related tweets in 2020?

1. Sub-question 1: Are there **communities** forming within these Twitter conversations, and what are they?
2. Sub-question 2: Are there **central actors** in the vaccine conversations during this year that drove the discussions towards certain topics?
3. Sub-question 3: What is the **sentiment** of the public with regard to vaccines and the virus in general?

Data Sources

In this research we tapped into several data sources with the intent to form a more complete view of the analysis given a few limitations on the individual datasets found. Our motive is to be able to have each data source compensate for each other's limitation and together form a more comprehensive view to answer our research questions above. These data sources are the following:

Covid Vaccine Tweets dataset (Kaggle)

The dataset contains over 200,000 tweets collected from 2020-02-12 to 2020-10-22. Data is collected daily by scraping tweets with hashtags related to coronavirus and the covid vaccines. Each record includes tweet id, tweet publish date, twitter account, tweet content and other related information. The data dictionary can be found in the appendix.

Source : https://www.kaggle.com/ritesh2000/covid19-vaccine-tweets

Limitation: Data dictionary of this data is not provided, the team had to make our own version by reading and understand the content of every column. Also some important information for a network is not available like retweet and user profile.

Mendeley Data by Norman Aguilar-Gallegos.

The dataset contains several databases related to Twitter posts on the coronavirus. The total data set contains 8,982,694 Twitter posts (tweets) which was extracted by finding keywords related to 'coronavirus' (Aguilar-Gallegos, 2020).

Source: https://data.mendeley.com/datasets/7ph4nx8hnc/1.

Limitation: The dataset is sizable, but the period covered is only between January 21 and February 12, 2020.

Twitter API

The Twitter API is a programming interface which allows developers to access twitter-account or tweet-related information. We are able to use this platform to be able to extract tweets ourselves when we need a higher quantity of tweets for a certain period.

Source : https://developer.twitter.com/en/docs/twitter-api

Limitation: Standard search API is free but restricted to search in past 7 days of tweets only and it allows 300 calls per 15 minutes. Our team also cannot perform scraping to get the Covid-Vaccine tweets data in 2021 to enrich our dataset, as only 2020 is currently available.

Network Definition

By using the Twitter dataset, we are presented with several options on how to define the nodes and edges of our graph analyses, as there are various methods by which Twitter users can interact with one another - including following, replying to, mentioning, and retweeting.

Amongst these, we identified retweeting as the priority connection to take a look at – as retweeting is the action we consider to be the most potent among the Twitter interactions above. This is simply because retweeting carries the same message of the original tweeter and pushes that message to the retweeter's followers in its raw form, unlike replying or mentioning where this is not necessarily the case. If thousands retweet *therealdonaldtrump'*s tweet for example, then his message is suddenly broadcasted to tens of thousands more accounts.

We will however go into the other Twitter interactions as well, the nodes and edges of which we define here:

1. Follow
    - Nodes: users
    - Edges: user ---- [follows] ---> user
2. Reply to
    - Nodes: users
    - Edges: user ---- [replies to] ---> user
3. Mention
    - Nodes: users
    - Edges: user ---- [mentions] ---> user
4. Retweet
    - Nodes: users
    - Edges: user ---- [retweets] ---> user

Exploratory Data Analysis (EDA)

Before we go into the network analysis and insights, we first wanted to get a good grasp of the data we have onhand.

The following two graphs illustrate the active users on the Twitter timeline during 2020. We have used the Kaggle 200k tweets dataset here as that is the data source that covers the whole year. The graphics represent the users who received a huge number of re-tweets and mentions in their posts.

Top retweets

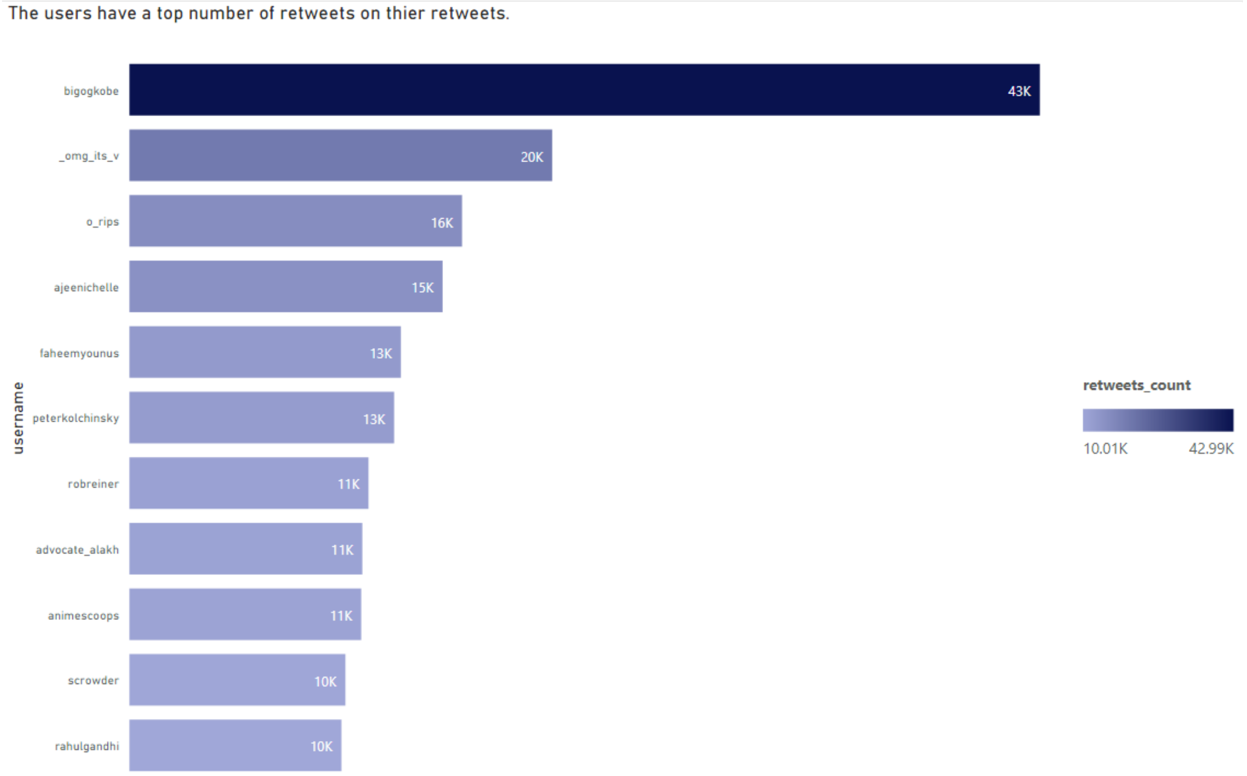


Figure 1. Top Users by Retweets

Replies

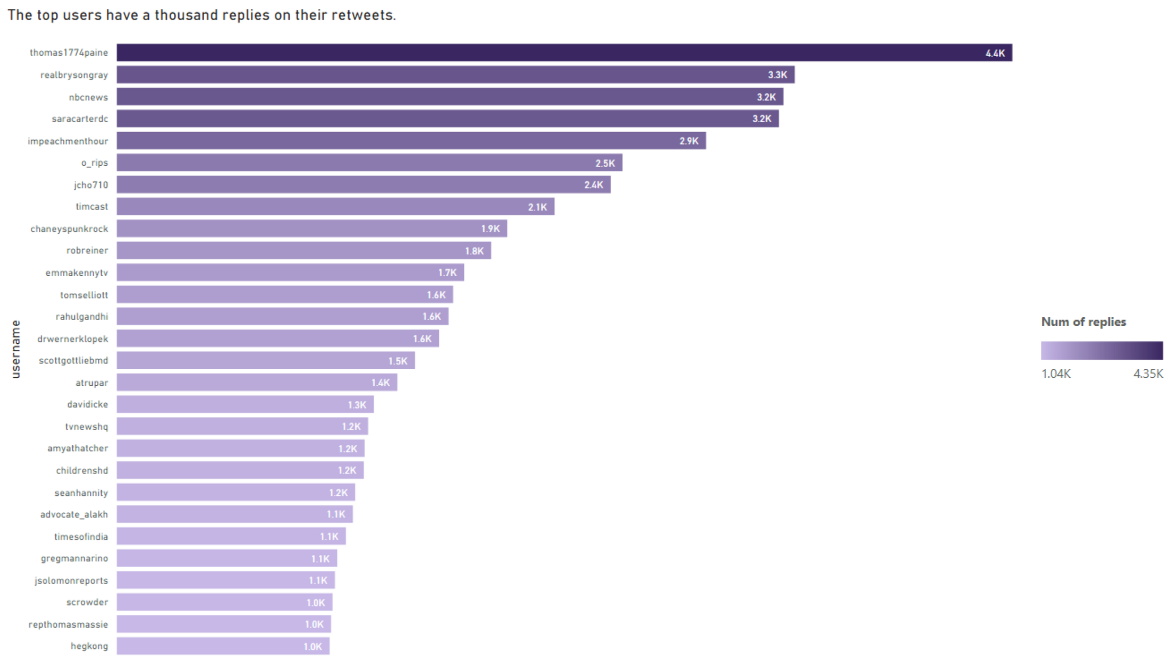
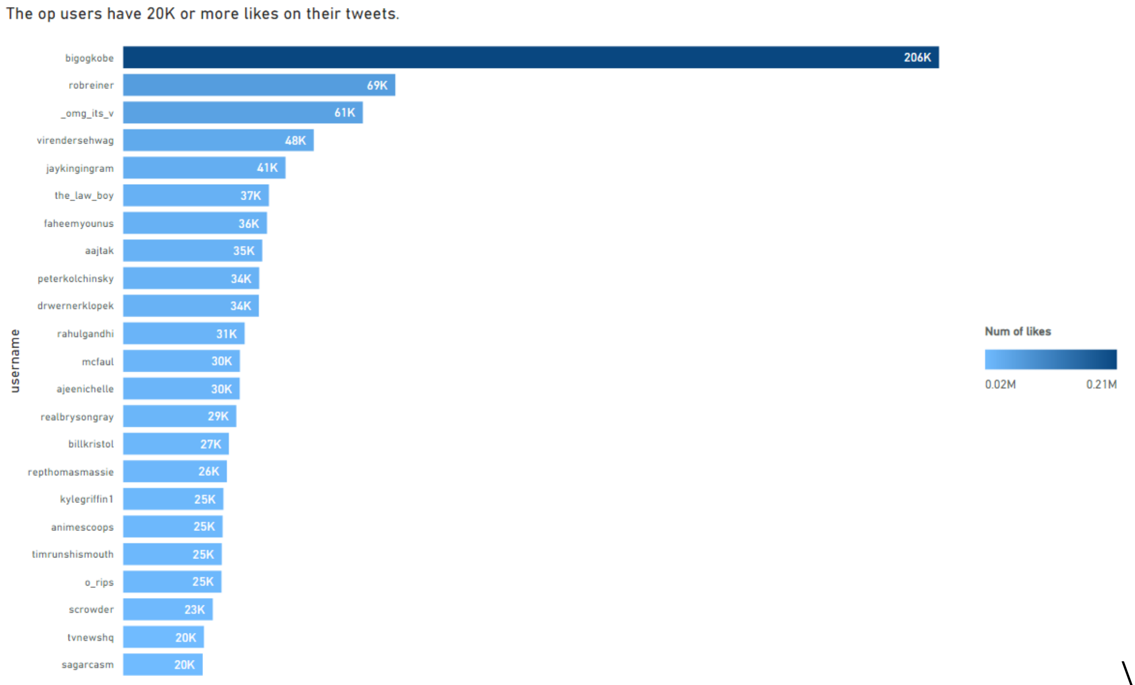


Figure 2. Top Users by Replies

Top likes

The favourite tweets from a specific (public) Twitter account are shown as "likes". The figure below shows the most popular users whose tweets have received 20,000 and more likes.



\

Figure 3. Top Users by Likes

According to the main results, the anonymous Twitter account "@bigogkobe" ranked first, while the filmmaker Rob Reiner "@robreiner" has their tweet coming in at second. This provides an initial insight that the most liked or retweeted Twitter accounts can either be known personalities or even just regular accounts, which we will explore further in the upcoming analyses.

Wordcloud

A word cloud is "an image of words used in a particular text or topic, and the size of each word indicates its frequency or importance." Based on the figure below, we can find out that the most popular word during the release period was related to Covid and vaccine. We also see how conversations revolved around the various suppliers of the vaccine: such as moderna, astrazeneca, covaxin, pfizerbiontech, and others.

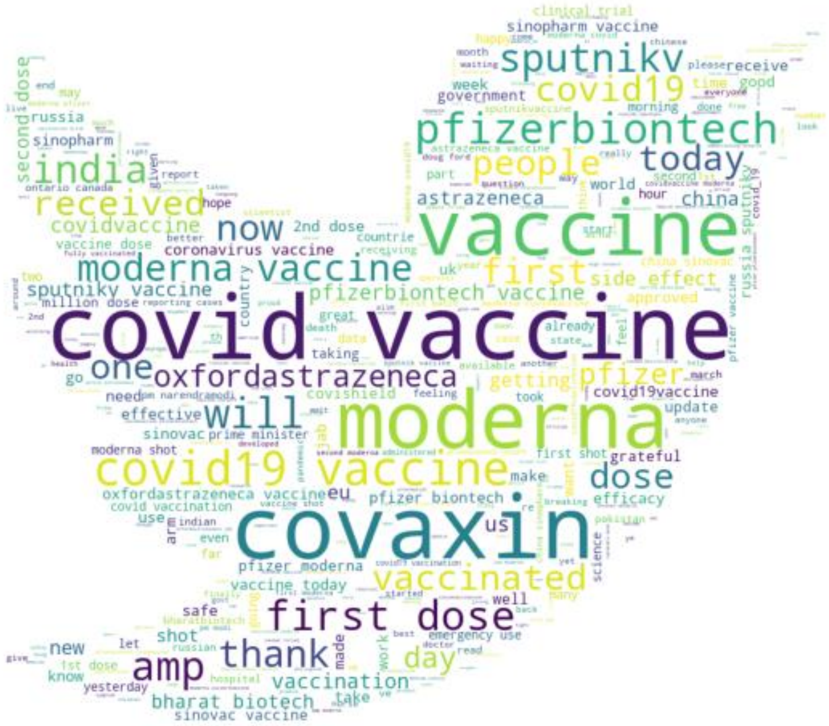


Figure 4. Wordcloud of Tweets, 2020

Hashtag Analysis

A relevant feature of Twitter that can act as a great categorical tool is the use of hashtags. It has been common practice among Twitter users to add these hashtags at the end of their tweets to signify the topic of what they wrote, and so by summarising this aspect of the data, we’re able to gain a few insights regarding the shape of the conversation of vaccine-related tweets in the 2020 dataset.

In Figure 5 below which outlines the top hashtags used during the period, we see that users have generally adopted #covidvaccine the most when talking about the vaccines in Twitter. A few unique highlights from this list would be #russia and #russianvaccine, which gained news mileage from their announcement of the Sputnik vaccine sometime August last year (Khurshudyan & Johnson, 2020). #covid19india, #trump, and #billgates are also specific hashtags that divert from the more general #vaccine or #corona hashtags which would be an interest for us to zoom in on.

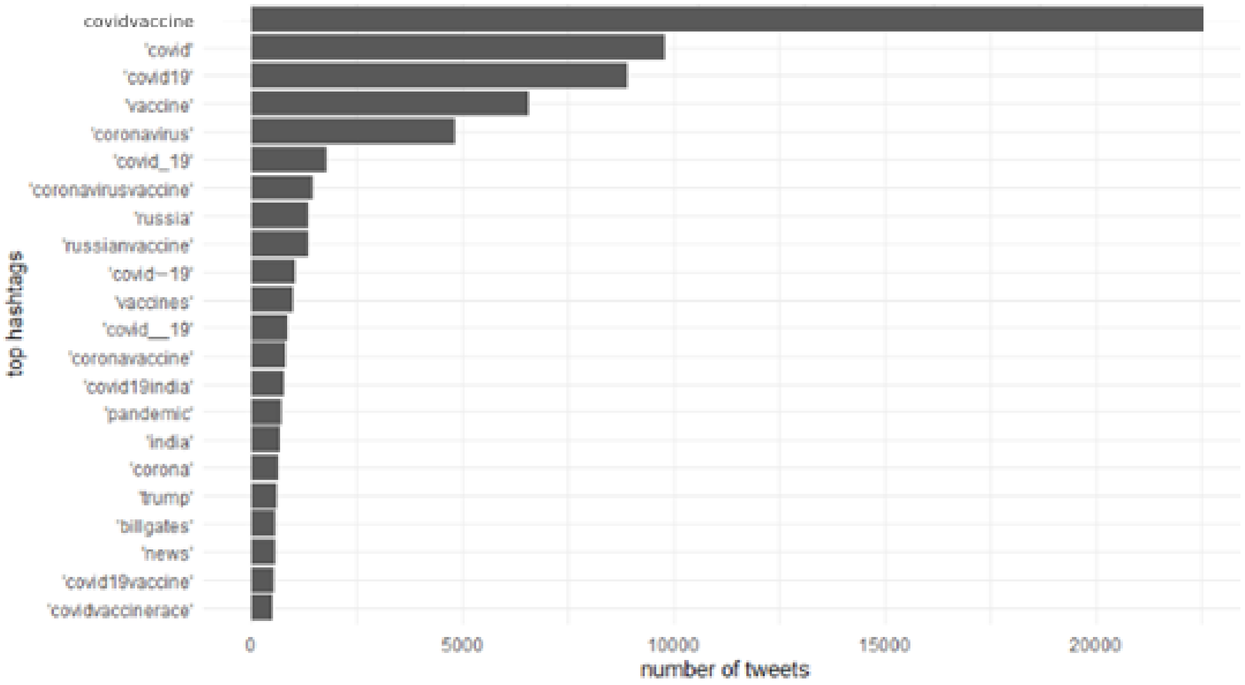


Figure 5. Top Hashtags by Number of Tweets

By plotting the movement of these hashtags over the days as seen in Figure 6, we immediately observe how the date of August 11, 2020 has an incredible spike of tweets in a single day. This matches exactly with Russia’s announcement of their Sputnik vaccine, which was the first vaccine formulation announced to be approved by a government. Interestingly, none of the vaccine formulations such as Moderna or Astrazeneca made it to the top hashtags list in Figure 5, so in some respects, Russia has been successful in getting public interest to the very first vaccine that was approved: Sputnik.

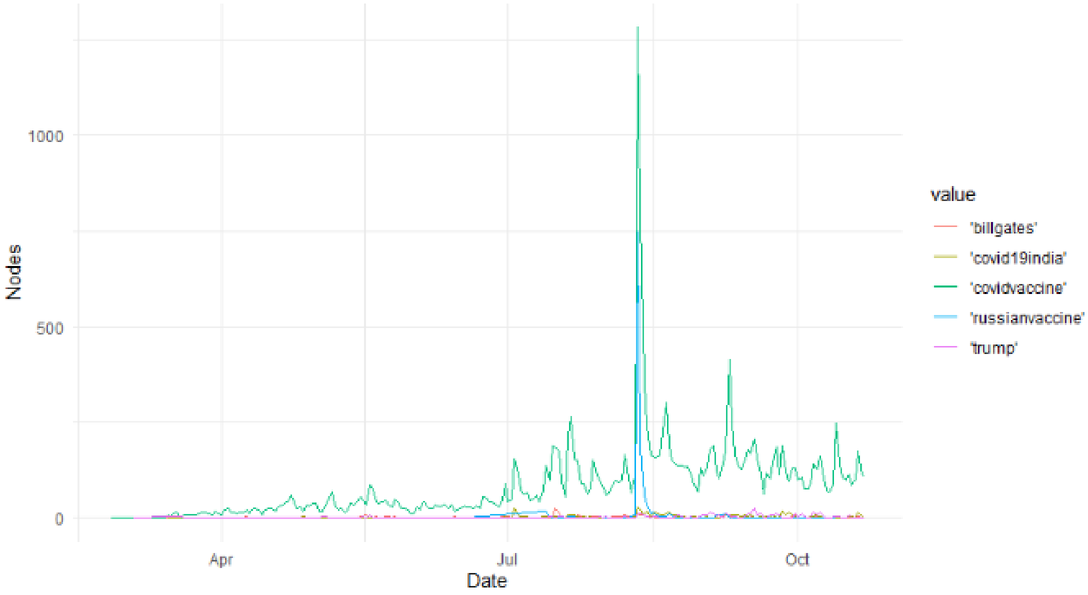
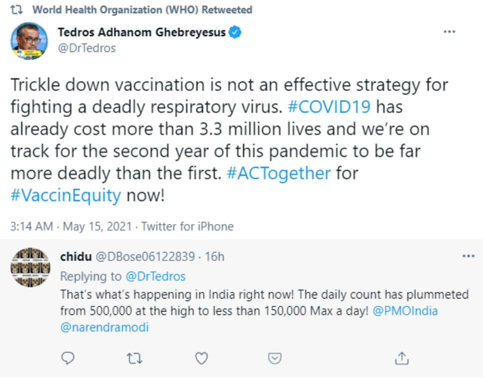


Figure 6. Use of Selected Hashtags by Date

Network Visualization

Visualising the network allows us to simplify the complex relationship between entities, thus helping us to identify the pattern, understand the context, and gather insights. Therefore, we represent the Twitter datasets in a directed graph, where the user is the node, and the interaction between them is the edges. For instance, when user A replies or retweets user's B tweet or when user A mentions user's B in the tweet then there is an edge from node A towards node B.



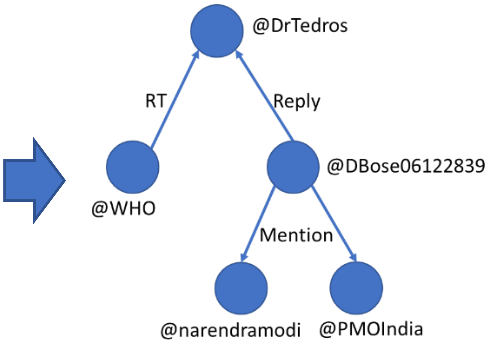
World Health Organization (WHO) Retweeted

Tedros Adhanom Ghebreyesus
@DrTedros

Trickle down vaccination is not an effective strategy for fighting a deadly respiratory virus. #COVID19 has already cost more than 3.3 million lives and we're on track for the second year of this pandemic to be far more deadly than the first. #ACTogether for #VaccinEquity now!

3:14 AM · May 15, 2021 · Twitter for iPhone

chidu @DBose06122839 · 16h
Replying to @DrTedros
That's what's happening in India right now! The daily count has plummeted from 500,000 at the high to less than 150,000 Max a day! @PMOIndia @narendramodi



The Kaggle dataset has provided several attributes to construct the network, such as mentions, reply to, and user retweet. Based on the ± 210.000 tweets in the dataset, we generate a network that consists of 133.658 nodes and 158.360 edges. However, all the edges found are the 'mention' or 'reply-to' relationship, there are no 'retweet' interactions found in the dataset.

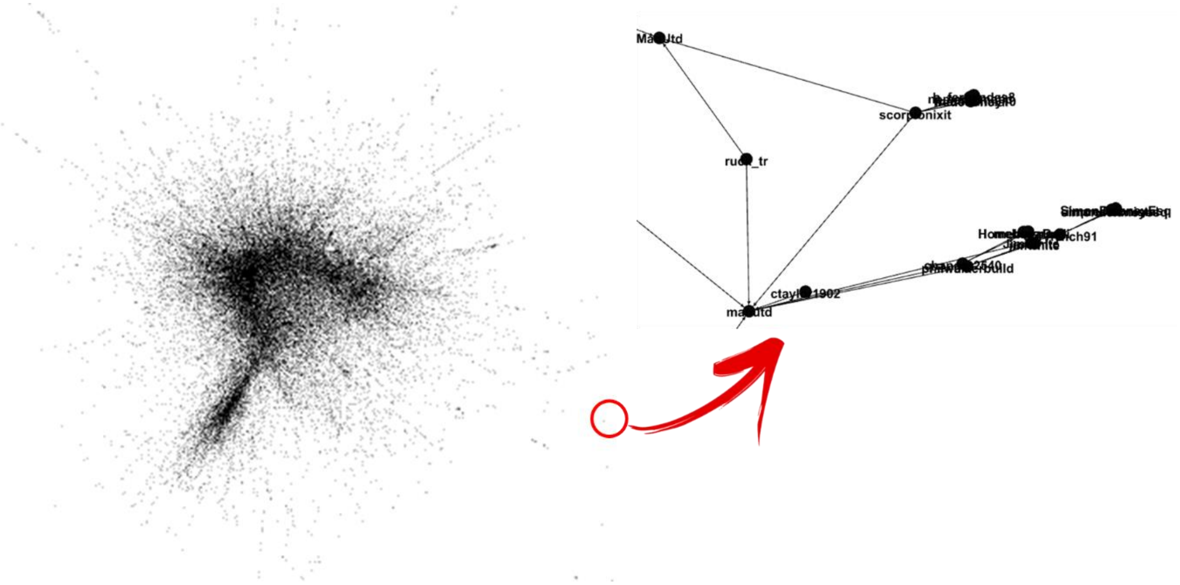


Figure 7. Kaggle Tweets Dataset Network Visualisation

As retweet is the primary interaction that we want to explore, we look into an alternative dataset collected by Norman Aguilar-Gallegos and uploaded at Mendeley.com (Aguilar-Gallegos, 2020). This dataset consists of 8,982,694 tweets collected from 21 January to 12 February 2020 (23 days) based on the keyword “Coronavirus,” so this dataset is not only about vaccines but coronavirus in general. However, due to computation power limitation, we use the tweets in English and come from the last two days only (11 – 12 February) that comprises 582,833 tweets.

We then filtered the interactions provided in the dataset so that it includes the selected tweets only. Initially, we obtain 529,608 interactions that consist of:

a. 429,834 retweets
b. 55,298 reply-to
c. 44,476 mentions

After combining the parallel edges (edges connecting the same pair of nodes), we got 461,074 edges connecting 285,553 nodes.

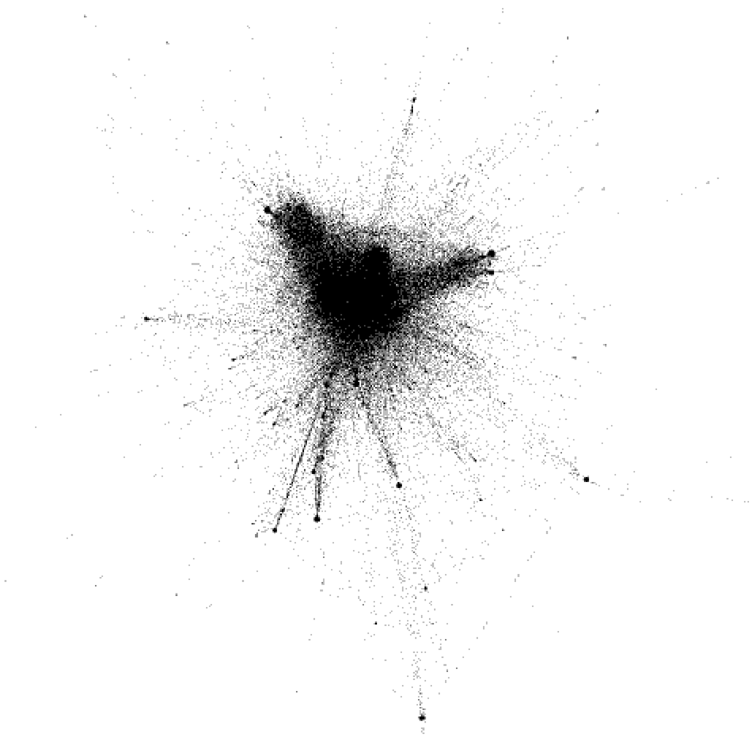


Figure 8. Mendeley Dataset Network Visualisation

One of the main challenges in analyzing the latter dataset is the absence of the tweet's text where the dataset only provides the status/tweet id. So, we utilize a python-based library called Hydrator - which encapsulates the Twitter API - to collect the text and other relevant attributes then store the tweets into a JSON file and the interactions in a CSV file. Hydrator also ensures that the use of the Twitter API complies with Twitter policy, such as the maximum number of tweets that can be collected within a particular period.

With this overview of how we tackled network visualization along with some of our limitations, we now go into exploring our research questions highlighted at the beginning of the paper.

Community Detection

Community Detection aims to identify clusters of nodes where the nodes within a cluster are more similar compared to the other nodes outside its group. According to Girvan and Newman, network nodes are tightly connected in knit groups within communities and loosely connected between communities (Girvan and Newman, 2002). Community detection can help us to identify people with a similar interest in a social network. In our case, it is interesting to see whether the false information around COVID propagates in particular groups.

We are using Gephi to discover and visualize the community. The modularity module in Gephi based Louvain's algorithm where the Modularity Q is defined as:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j),$$

A_{ij} is the weight of the edge between i and j.
k_i is the sum of the node's weight attached to the node i (degree of node i).
c_i is the community to which node i is assigned.
$\delta(c_i, c_j)$ is Kronecker delta function. $\delta(c_i, c_j)$ = 1 if c_i=c_j, 0 otherwise.
$m = \frac{1}{2} \sum A_{ij}$

Vaccine Dataset (Kaggle)

There are 19.269 communities found on this dataset. However, only three of them have a substantial number of nodes (> 5.000 nodes).

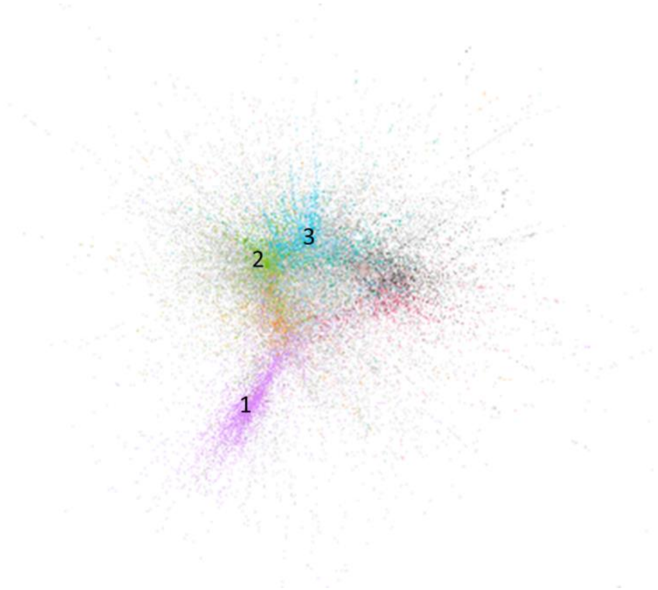
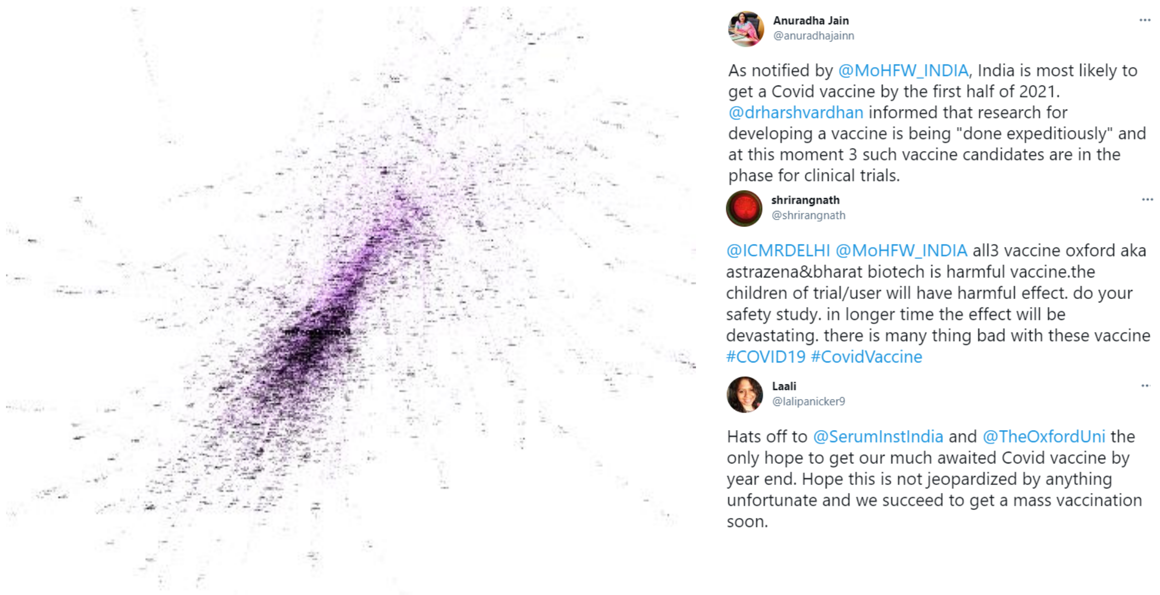


Figure 9. Kaggle Dataset Community Visualisation

The first community consists of 11566 nodes. We found that @narendramodi (India Prime Minister), @drharshvardhan (Minister of Health - India), @icmrdelhi (Indian Council of Medical Research), and @SerumInstIndia (a vaccine manufacturer) are the nodes with the highest degree in this cluster. The tweets in this group range from news, support the vaccine development, and anti-vaccine tweets. However, most of the tweets are from India.



Anuradha Jain
@anuradhajainn

As notified by @MoHFW_INDIA, India is most likely to get a Covid vaccine by the first half of 2021. @drharshvardhan informed that research for developing a vaccine is being "done expeditiously" and at this moment 3 such vaccine candidates are in the phase for clinical trials.

shrirangnath
@shrirangnath

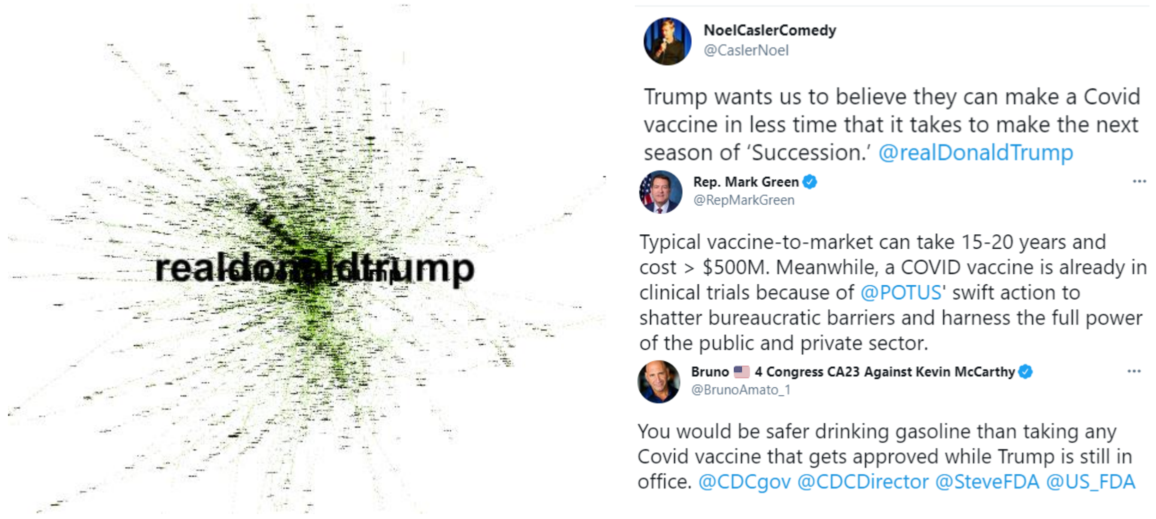
@ICMRDELHI @MoHFW_INDIA all3 vaccine oxford aka astrazena&bharat biotech is harmful vaccine.the children of trial/user will have harmful effect. do your safety study. in longer time the effect will be devastating. there is many thing bad with these vaccine #COVID19 #CovidVaccine

Laali
@lalipanicker9

Hats off to @SerumInstIndia and @TheOxfordUni the only hope to get our much awaited Covid vaccine by year end. Hope this is not jeopardized by anything unfortunate and we succeed to get a mass vaccination soon.

Figure 10. Kaggle Dataset - Community 1

Looking at the second community that consist of 8047 nodes, we found that @realdonaltrump (45th President – US), @POTUS, @WhiteHouse, and @SteveFDA (FDA commissioner) are the nodes with the largest degree. The tweets in this group also vary, for instance, news, supporting the Trump administration policy, or against it. Interestingly, despite having the highest degree in the community, @realdonaldtrump only has the inward edges, without outward edges (posting any tweets).



NoelCaslerComedy
@CaslerNoel

Trump wants us to believe they can make a Covid vaccine in less time that it takes to make the next season of 'Succession.' @realDonaldTrump

Rep. Mark Green
@RepMarkGreen

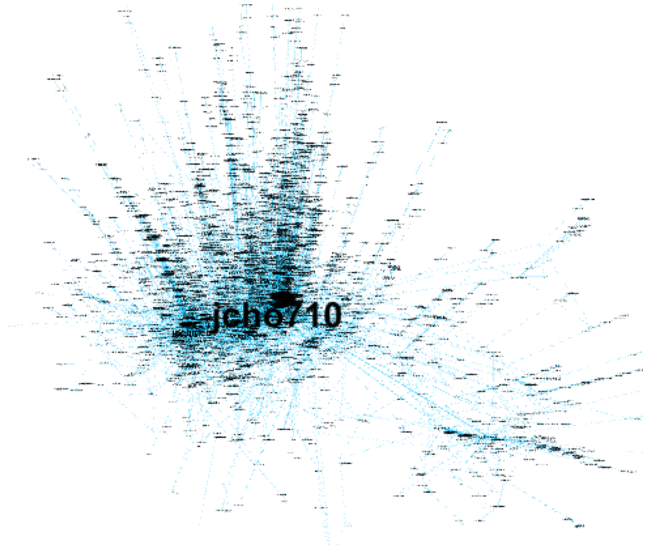
Typical vaccine-to-market can take 15-20 years and cost > $500M. Meanwhile, a COVID vaccine is already in clinical trials because of @POTUS' swift action to shatter bureaucratic barriers and harness the full power of the public and private sector.

Bruno 🇺🇸 4 Congress CA23 Against Kevin McCarthy
@BrunoAmato_1

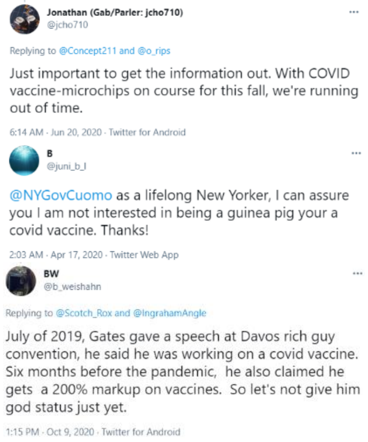
You would be safer drinking gasoline than taking any Covid vaccine that gets approved while Trump is still in office. @CDCgov @CDCDirector @SteveFDA @US_FDA

Figure 11. Kaggle Dataset - Community 2

The third community consists of 6926 nodes. We found that @jcho710, @nygovcuomo (New York State Governor – US), and @joebiden (47th US President) are the nodes with the largest degree. Similar with the other cluster we found no particular topic of interest in this group, despite the fact that the user @jcho710 that with the highest degree in this community spreading several false information about covid vaccine.



Tweets:


Jonathan (Gab/Parler: jcho710)
@jcho710

Replying to @Concept211 and @o_rips

Just important to get the information out. With COVID vaccine-microchips on course for this fall, we're running out of time.

6:14 AM · Jun 20, 2020 · Twitter for Android

B
@juni_b_l

@NYGovCuomo as a lifelong New Yorker, I can assure you I am not interested in being a guinea pig your a covid vaccine. Thanks!

2:03 AM · Apr 17, 2020 · Twitter Web App

BW
@b_weishahn

Replying to @Scotch_Rox and @IngrahamAngle

July of 2019, Gates gave a speech at Davos rich guy convention, he said he was working on a covid vaccine. Six months before the pandemic, he also claimed he gets a 200% markup on vaccines. So let's not give him god status just yet.

1:15 PM · Oct 9, 2020 · Twitter for Android

Figure 12. Kaggle Dataset - Community 3

COVID Dataset (Mendeley Data)

There are 6486 communities found on this dataset. However, only four of them with a substantial number of nodes or at least 5% from the total nodes.

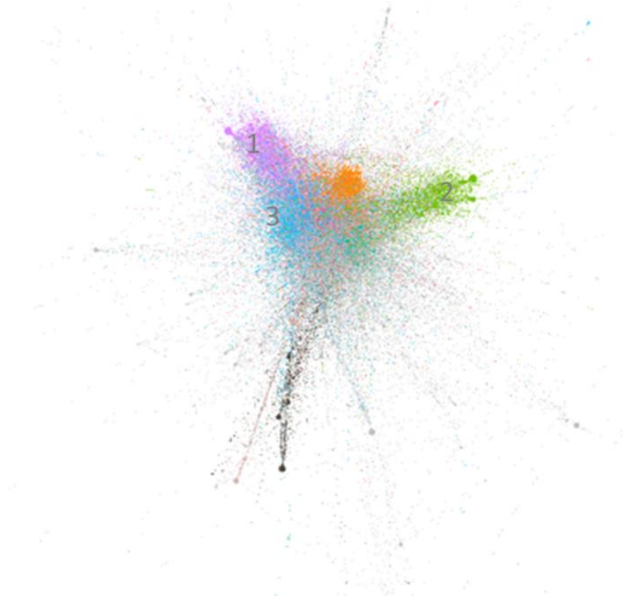
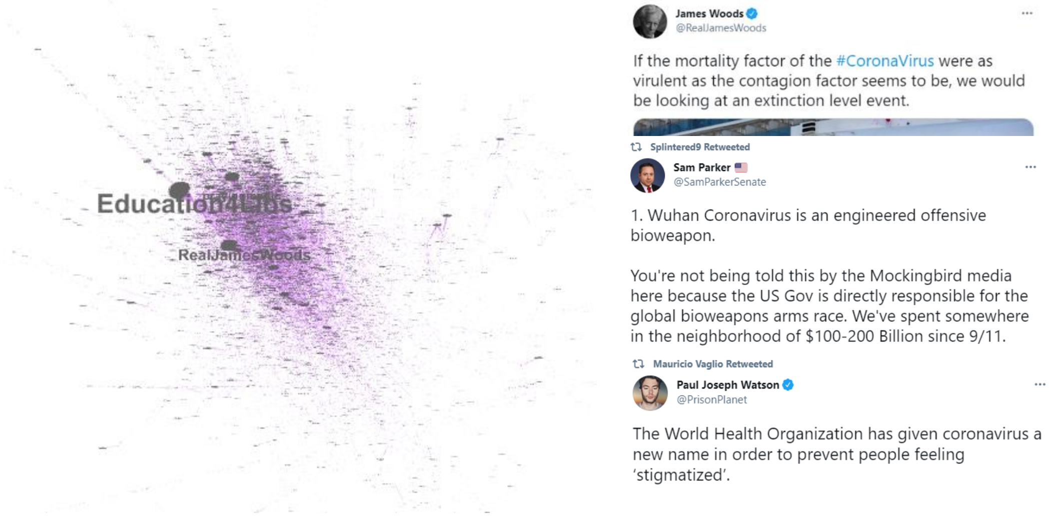


Figure 13. Mendeley Dataset Community Visualisation

Looking at the first community that consists of 17,921 nodes, we found that @Education4Libs, @prayingmedic (both of these accounts have been suspended), @RealJamesWoods, and @PrisonPlanet are the nodes with the largest degree. We found that most of the users in this community are Republican or President Trump supporters. We also found several false information and conspiracy theories spreading in the community.



James Woods
@RealJamesWoods

If the mortality factor of the #CoronaVirus were as virulent as the contagion factor seems to be, we would be looking at an extinction level event.

Splintered9 Retweeted
Sam Parker 🇺🇸
@SamParkerSenate

1. Wuhan Coronavirus is an engineered offensive bioweapon.

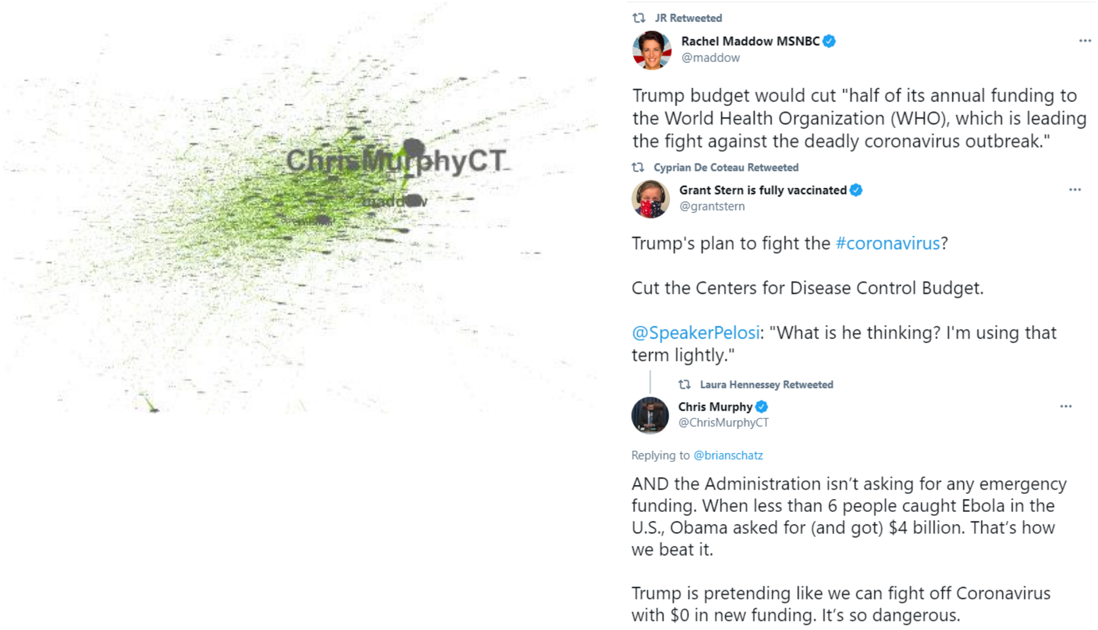
You're not being told this by the Mockingbird media here because the US Gov is directly responsible for the global bioweapons arms race. We've spent somewhere in the neighborhood of $100-200 Billion since 9/11.

Mauricio Vaglio Retweeted
Paul Joseph Watson
@PrisonPlanet

The World Health Organization has given coronavirus a new name in order to prevent people feeling 'stigmatized'.

Figure 14. Mendeley Dataset - Community 1

The second community consists of 14,583 nodes. We found that @ChrisMurphyCT (US Senator), @maddow (MSNBC host), @SpeakerPelosi (US House of Representative), @grantstern are the nodes with the largest degree. We found that most of the users in this community are Democrats or the critics of President Trump policy.



JR Retweeted
Rachel Maddow MSNBC
@maddow

Trump budget would cut "half of its annual funding to the World Health Organization (WHO), which is leading the fight against the deadly coronavirus outbreak."

Cyprian De Coteau Retweeted
Grant Stern is fully vaccinated
@grantstern

Trump's plan to fight the #coronavirus?

Cut the Centers for Disease Control Budget.

@SpeakerPelosi: "What is he thinking? I'm using that term lightly."

Laura Hennessey Retweeted
Chris Murphy
@ChrisMurphyCT

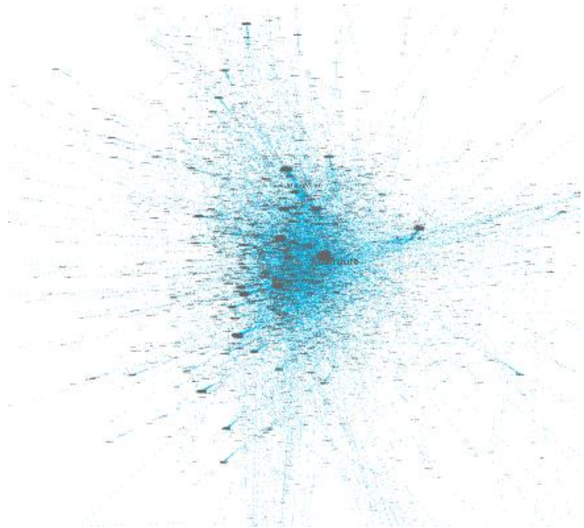
Replying to @brianschatz

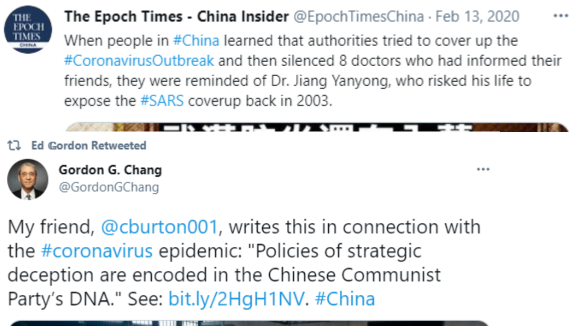
AND the Administration isn't asking for any emergency funding. When less than 6 people caught Ebola in the U.S., Obama asked for (and got) $4 billion. That's how we beat it.

Trump is pretending like we can fight off Coronavirus with $0 in new funding. It's so dangerous.

Figure 15. Mendeley Dataset - Community 2

The third community consists of 14,352 nodes. We found that @howroute, @IsChinar, @livecrisisnews, and @EpochTimesChina are the nodes with the largest degree. However, we can't retrieve the tweets from @howroute, @IsChinar, and @livecrisisnews as all the accounts have been suspended. Most of the tweets are talking about the opposition to China's government action in handling the pandemic.





The Epoch Times - China Insider @EpochTimesChina · Feb 13, 2020 ···

When people in #China learned that authorities tried to cover up the #CoronavirusOutbreak and then silenced 8 doctors who had informed their friends, they were reminded of Dr. Jiang Yanyong, who risked his life to expose the #SARS coverup back in 2003.

Ed Gordon Retweeted

Gordon G. Chang ···
@GordonGChang

My friend, @cburton001, writes this in connection with the #coronavirus epidemic: "Policies of strategic deception are encoded in the Chinese Communist Party's DNA." See: bit.ly/2HgH1NV. #China

Figure 16. Mendeley Dataset - Community 3

Community detection can group similar users/nodes based on their interest in a particular topic. We found that it's hard to distinguish the communities from the Kaggle dataset, which is more apparent on the coronavirus dataset. We assume that it's caused by the difference of interaction type available between the vaccine (Kaggle) and the coronavirus dataset, where the retweet interaction is absent from the Kaggle dataset. The user retweets means that the user high likely agrees on the tweet's content while mentions or reply-to convey both stances, either degree or disagree.

What's apparent from this exercise is the existence and prevalence of an 'echo chamber' phenomenon. Community clusters in social media tend to only see and hear information that they already agree with – thus only furthering the distance between them versus those that oppose their views. This is best evidenced by the separate communities of Democrats and Republicans as seen in the networks above. These findings enrich the discussion on the benefits versus the costs of the existence of social media platforms like Twitter – where some argue that these channels have made connections closer, findings like in the analysis above also show another side of the story.

Centrality of Tweets

Centrality Coefficients

Importance of a twitter account can be observed by the attention of tweets of these accounts. In this part, the centrality of twitter accounts is measured in a Twitter Reply Network. Centralities of each twitter account are calculated by using library “networkx” in Python.

Following is the sample of edge list and top 20 centrality of twitter accounts.

username	reply_to_username	Weight
jurg_ames	realDonaldTrump	186
jcho710	ChuckCallesto	139
jcho710	JoeBiden	163

Edge List sample

Rank	username	centrality	Type
1	realdonaldtrump	0.030653	Government / Organization
2	youtube	0.011926	Company
3	jcho710	0.011118	Others / Individual
4	narendramodi	0.005903	Government / Organization
5	joebiden	0.005290	Government / Organization
6	who	0.004826	Government / Organization
7	pmoindia	0.004362	Government / Organization
8	cnn	0.004108	Media
9	mailonline	0.003696	Media
10	potus	0.003434	Government / Organization
11	billgates	0.003359	Celebrity / Public Figure
12	matthancock	0.002641	Government / Organization
13	whitehouse	0.002626	Government / Organization
14	mohfw_india	0.002394	Government / Organization
15	nytimes	0.002364	Media
16	kamalaharris	0.002342	Government / Organization
17	us_fda	0.002230	Government / Organization
18	drharshvardhan	0.002087	Government / Organization
19	cdcgov	0.002050	Government / Organization
20	nygovcuomo	0.001968	Government / Organization

Top 20 centrality of twitter accounts

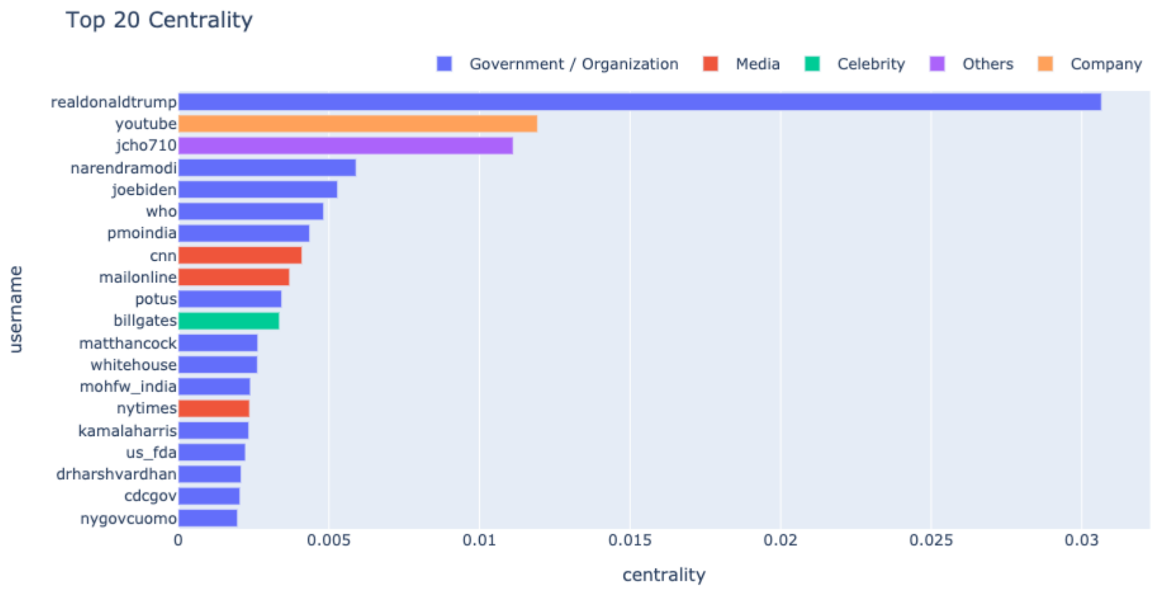
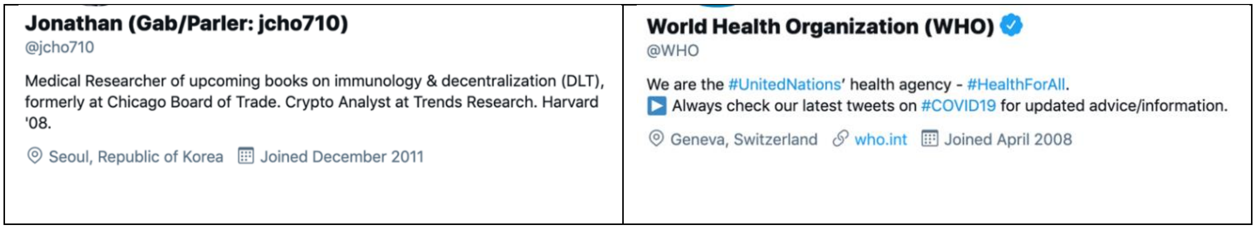


Figure 17. Top 20 Most Central Accounts

The result shows that Donald Trump is the most important in terms of centrality of covid vaccine tweet and he's centrality coefficient is far higher than other accounts. Also, the results indicates that the tweets from Government or International Organization related parties are more importance comparing to other parties like media or celebrities regarding the information spreading by interaction between accounts.

An interesting finding is that there is an anonymous Twitter account "jcho710" ranked at 3rd regarding centrality. Unlike other Twitter accounts in top 20 list, this account does not have any blue verified badge, which means this not an account of public interest nor authentic (Twitter, 2021).



Jonathan (Gab/Parler: jcho710)
@jcho710

Medical Researcher of upcoming books on immunology & decentralization (DLT), formerly at Chicago Board of Trade. Crypto Analyst at Trends Research. Harvard '08.

Seoul, Republic of Korea Joined December 2011

World Health Organization (WHO) ✓
@WHO

We are the #UnitedNations' health agency - #HealthForAll.
▶ Always check our latest tweets on #COVID19 for updated advice/information.

Geneva, Switzerland who.int Joined April 2008

Account jcho710 (left) doesn't have blue verified badge while WHO has (right)

According to user profile of this account, he is a self-claimed medical researcher. Further investigation is done on the number of mentions and reply, this account is the top accounts mentioned other accounts in his tweets and 2nd top accounts replied by other accounts. That's why the centrality of this account is high.

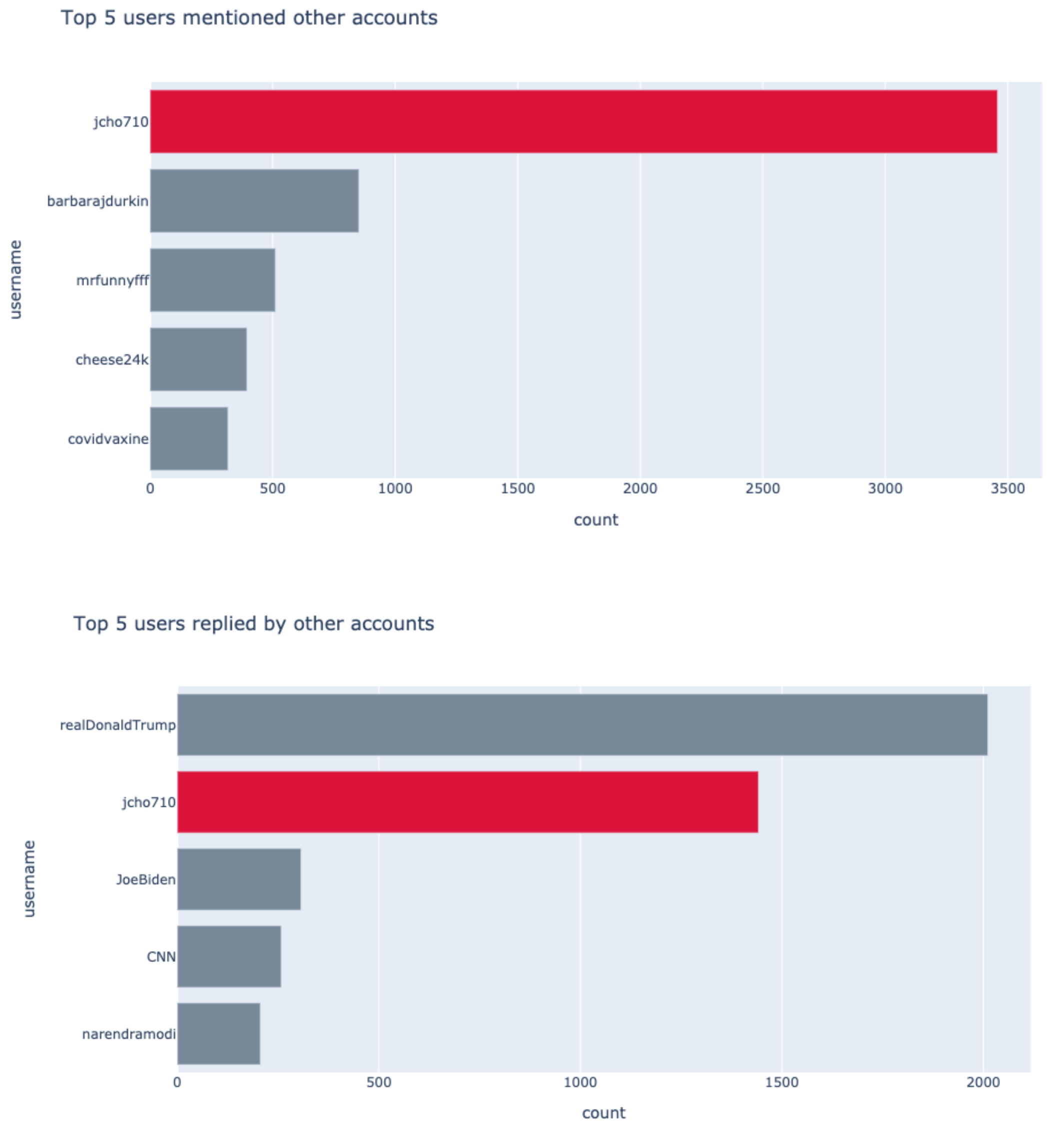


Figure 18. Top 5 Users Mentioned and Replied to by Other Accounts

This situation is quite concerning when talking about spreading of fake news if an anonymous account become a key information spreader on important topics like Covid.

K-Cores

The graph below shows the cored network with k=7. The number of nodes are 727. A dominate red dot is “realDonaldTrump”, who remains highest degrees in the network, which aligns with the centrality and indicate that this accounts is the most influential twitter account on topic about Covid-Vaccine in 2020 regarding information spreading by interaction of accounts.

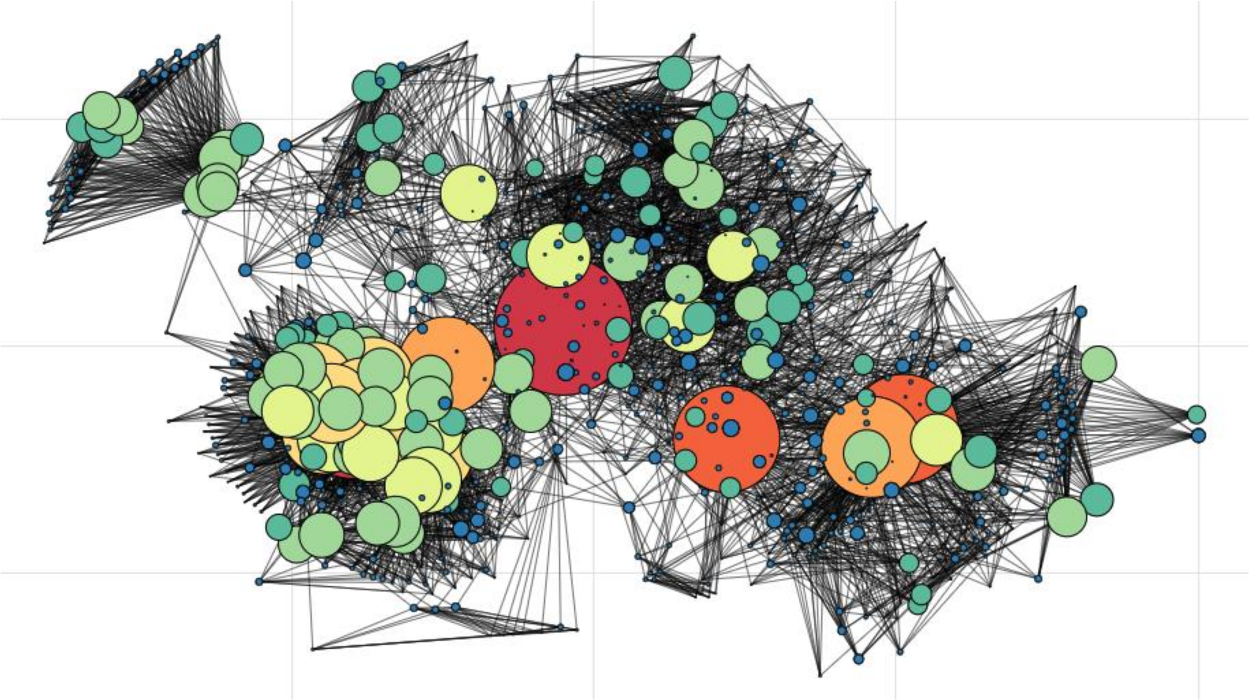


Figure 19. Cored network with k=7

Emotion and Sentiment Analysis

Emotion is overly complicated, multidimensional characteristic which reflects the personality and behavioral traits of humans (Sailunaz & Alhajj, 2019). In their daily life, people express their emotions on different issues, events, persons, environment and even every little thing. Despite the advancement in technology allows users of social networking platforms to demonstrate their emotions by audio and videos, text is still the most usual form of communication in social network.

Twitter data is a popular choice for text analysis tasks because of the limited number of characters 140 allowed and global use of Twitter to express opinion on different issues among people of all ages, races, cultures, and genders. In this project, we analyzed a twitter network for emotion analysis. We analyzed the emotions from tweets and their replies and formed an emotion network based on the text posted by users. From the emotion network, the influential people are expected to identify.

Kaggle Data:

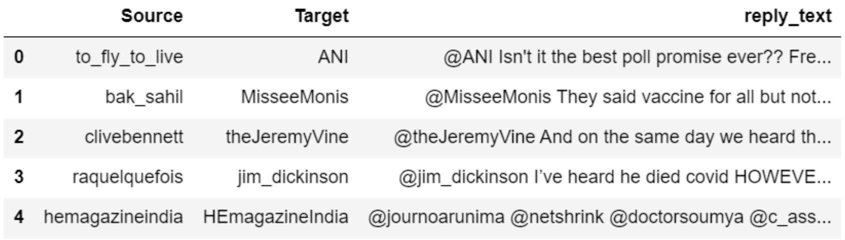
In this analysis, we used data available in Kaggle as mentioned in the Data Source session. The original dataset contains 209,929 records. We scoped our analysis to text in English language only. The filtered dataset consisted of 201,030 records out of which 51,358 users replied to 40,431 users. The total edges formed are 69,736 in the dataset.

The total nodes are 91789. As the original data is clean and structured, the data extraction was performed smoothly.

The limitation in the dataset for this emotion analysis, the tweets were scraped randomly, not based on specific of tweets. The column “conversation_id” which is considered as original tweet_id where original post and its replies are related to, is found to contain independent conversations or tweet_ids in the dataset. Altogether 188,437 conversation_ids are identified out of 201,030 observations.

Therefore, it can be concluded that all the relations are not tweet specific although the interaction of edges are formed between username who made the post and replied_to_username who was the original poster of the main tweet regardless of the tweet, under the main filtered topic of “#covid tweet”.

Network Data Formation:



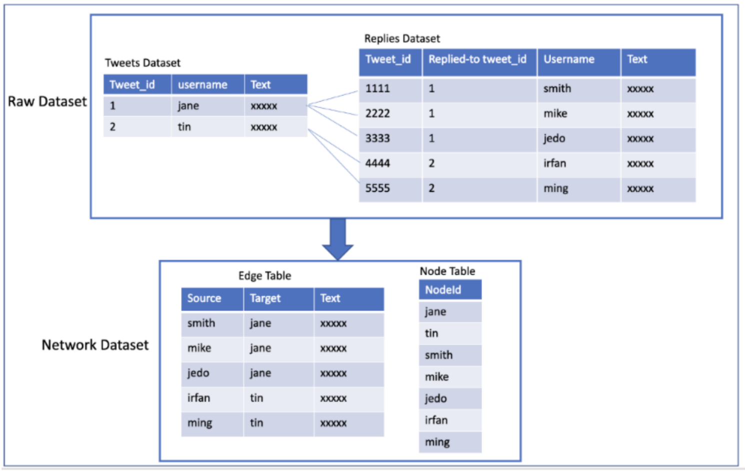
	Source	Target	reply_text
0	to_fly_to_live	ANI	@ANI Isn't it the best poll promise ever?? Fre...
1	bak_sahil	MisseeMonis	@MisseeMonis They said vaccine for all but not...
2	clivebennett	theJeremyVine	@theJeremyVine And on the same day we heard th...
3	raquelquefois	jim_dickinson	@jim_dickinson I’ve heard he died covid HOWEVE...
4	hemagazineindia	HEmagazineIndia	@journoarunima @netshrink @doctorsoumya @c_ass...

Twitter API Data:

Additionally, we scraped data from Twitter API to look at recent emotion trend in the small scale of data set. We applied Twitter API through the python package, tweepy. As COVID vaccine is the topic of the trend these days, we filtered the data using the keyword “#covidvaccine.

As the Twitter API allows only 300 tweets in 15 minutes within the same day and only past 7 days of the specified date to be requested. For the simplicity purpose, we scraped the data on the day of May 3 and 11. Firstly, the tweets by original poster were scraped. Next using the tweet_ids, the replies of each tweet_id were scraped individually. Below diagram illustrates how the raw collected tweet data had been transformed into the network dataset.

In this mini dataset, 1723 edges are formed with 937 source nodes to 15 Target nodes.



Raw Dataset

Tweets Dataset

Tweet_id	username	Text
1	jane	xxxxx
2	tin	xxxxx

Replies Dataset

Tweet_id	Replied-to tweet_id	Username	Text
1111	1	smith	xxxxx
2222	1	mike	xxxxx
3333	1	jedo	xxxxx
4444	2	irfan	xxxxx
5555	2	ming	xxxxx

Network Dataset

Edge Table

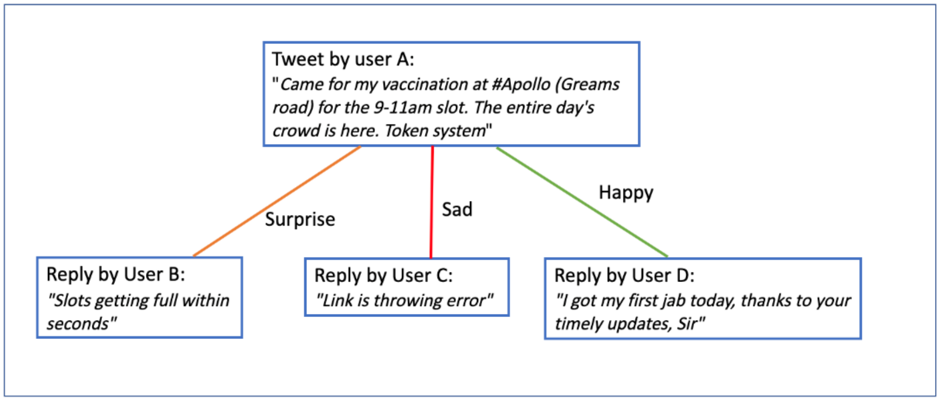
Source	Target	Text
smith	jane	xxxxx
mike	jane	xxxxx
jedo	jane	xxxxx
irfan	tin	xxxxx
ming	tin	xxxxx

Node Table

NodeId
jane
tin
smith
mike
jedo
irfan
ming

Node and Edges

In this analysis, the user accounts are the nodes. When a user replies to another user on its post, the edge is formed.



Tweet by user A:
"*Came for my vaccination at #Apollo (Greams road) for the 9-11am slot. The entire day's crowd is here. Token system*"

Reply by User B:
"Slots getting full within seconds"

Reply by User C:
"Link is throwing error"

Reply by User D:
"I got my first jab today, thanks to your timely updates, Sir"

Data Cleansing

Data pre-processing included cleaning the collected data and annotating the data according to sentiments and emotions. Tweet texts contain huge amount of unnecessary noise and symbols. The data cleansing process included as followed.

1) all user mentions were removed (I.e @john)

2) Hashtags were removed (I.e #vaccine)

3) web URLs were removed (I.e https://xxxxx)

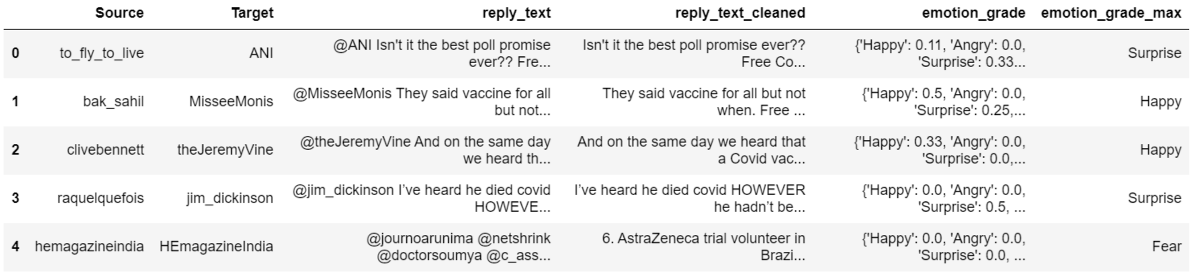
Emotion Detection

Finally, cleaned tweet replies were applied into the emotion detection algorithm. The observations were annotated with emotion results. In our analysis, ***text2emotion*** python library https://pypi.org/project/text2emotion/ is used, which list the rates in five basic human emotions – Happy, Angry, Surprise and Sad

for example - {‘Happy”: 1, “Angry”: 0.0, ‘Surprise’: 0.0, ‘Sad: 0.0, ‘Fear’: 0.0}.

When the emotion algorithm cannot detect the emotion from the text, the result is {‘Happy”: 0, “Angry”: 0, ‘Surprise’: 0, ‘Sad: 0.0, ‘Fear’: 0}.

In that case, the “neutral” is added manually.



	Source	Target	reply_text	reply_text_cleaned	emotion_grade	emotion_grade_max
0	to_fly_to_live	ANI	@ANI Isn't it the best poll promise ever?? Fre...	Isn't it the best poll promise ever?? Free Co...	{'Happy': 0.11, 'Angry': 0.0, 'Surprise': 0.33...	Surprise
1	bak_sahil	MisseeMonis	@MisseeMonis They said vaccine for all but not...	They said vaccine for all but not when. Free ...	{'Happy': 0.5, 'Angry': 0.0, 'Surprise': 0.25,...	Happy
2	clivebennett	theJeremyVine	@theJeremyVine And on the same day we heard th...	And on the same day we heard that a Covid vac...	{'Happy': 0.33, 'Angry': 0.0, 'Surprise': 0.0,...	Happy
3	raquelquefois	jim_dickinson	@jim_dickinson I've heard he died covid HOWEVE...	I've heard he died covid HOWEVER he hadn't be...	{'Happy': 0.0, 'Angry': 0.0, 'Surprise': 0.5, ...	Surprise
4	hemagazineindia	HEmagazineIndia	@journoarunima @netshrink @doctorsoumya @c_ass...	6. AstraZeneca trial volunteer in Brazi...	{'Happy': 0.0, 'Angry': 0.0, 'Surprise': 0.0, ...	Fear

Emotion Network Visualisation Using Kaggle Data

We used the software Gephi to visually present our emotion network.

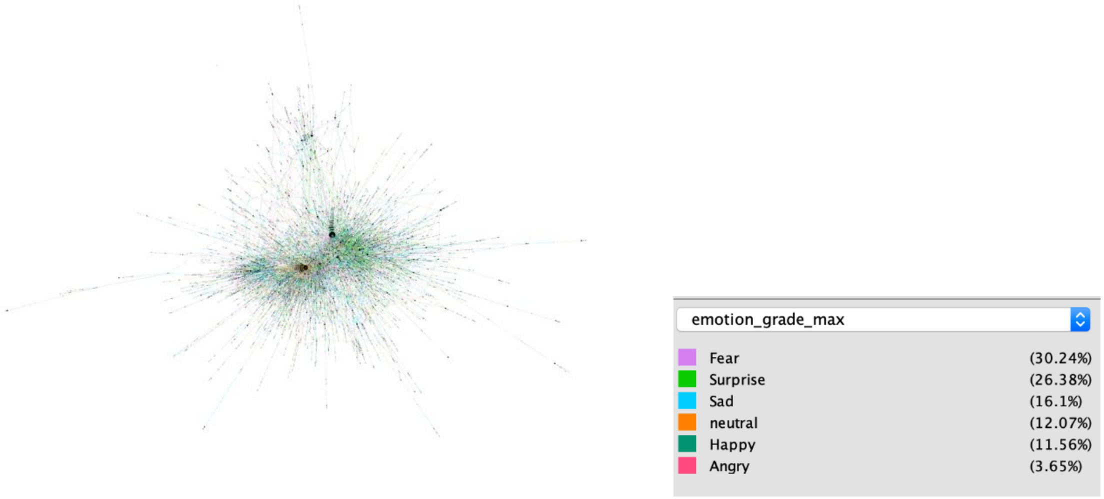
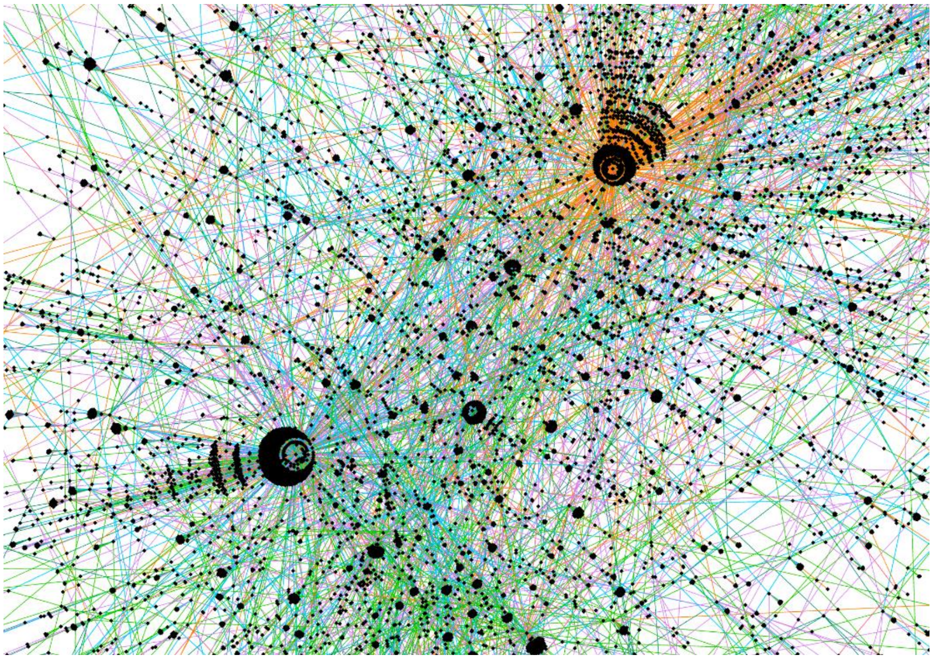
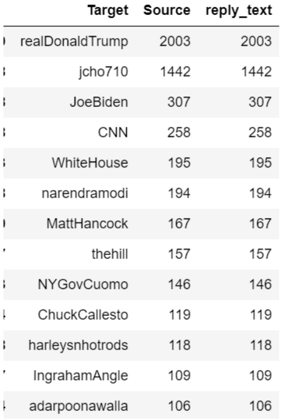


Figure 20. Sentiment Visualisation

From the overview, Fear and Surprise are seen as the biggest sentiments associated with the COVID-19 Tweets. It is to be noted however that the sentiments are more nuanced, as we also see other emotions such as Sad, Happy, and Angry.





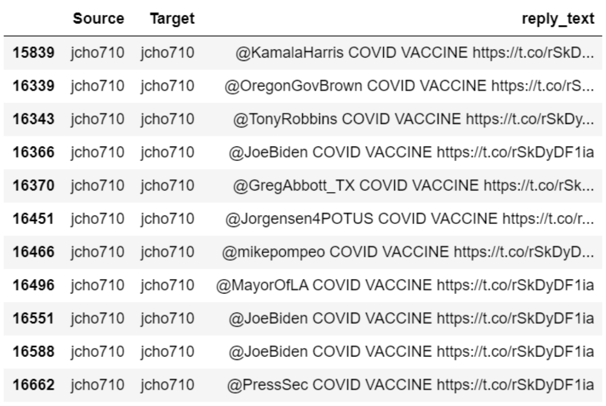
Target	Source	reply_text
realDonaldTrump	2003	2003
jcho710	1442	1442
JoeBiden	307	307
CNN	258	258
WhiteHouse	195	195
narendramodi	194	194
MattHancock	167	167
thehill	157	157
NYGovCuomo	146	146
ChuckCallesto	119	119
harleysnhotrods	118	118
IngrahamAngle	109	109
adarpoonawalla	106	106

Figure 21. Closer View of Sentiment Graph Visualisation

In the closer look, at the central of the network, the two largest nodes are identified. One node with neutral emotion and another node mixed with Fear, Surprise and Sad emotions. The dominant accounts with highest numbers of replies in descending order as described in the in the screen on the right.

The account @realDonaldTrump having 2003 replies which are surrounded by Fear (purple) and Surprise (green) when talking about Covid vaccination. The color does not look positive.

Another outstanding node is by the user @jcho710. The replies he received were 1442 which was more than the tweet replies received by JoeBiden and CNN. The major color of its edges are Neutral in orange. As the user @jcho710 is an individual twitter account. It drew our attention to examine further in the dataset. What we found out that is the user had been posting the link (https://t.co/rSkDyDF1ia) of his own post by tagging the government public figures. Out of 1442 edges in the dataset, 1432 edges are his self-interactions. The only 10 replies are made by other twitter users. The user seemed to be intentionally propagating the message to the high-profile politicians in the Unites States as shown in the below screen shot of data query.



	Source	Target	reply_text
15839	jcho710	jcho710	@KamalaHarris COVID VACCINE https://t.co/rSkD...
16339	jcho710	jcho710	@OregonGovBrown COVID VACCINE https://t.co/rS...
16343	jcho710	jcho710	@TonyRobbins COVID VACCINE https://t.co/rSkDy...
16366	jcho710	jcho710	@JoeBiden COVID VACCINE https://t.co/rSkDyDF1ia
16370	jcho710	jcho710	@GregAbbott_TX COVID VACCINE https://t.co/rSk...
16451	jcho710	jcho710	@Jorgensen4POTUS COVID VACCINE https://t.co/r...
16466	jcho710	jcho710	@mikepompeo COVID VACCINE https://t.co/rSkDyD...
16496	jcho710	jcho710	@MayorOfLA COVID VACCINE https://t.co/rSkDyDF1ia
16551	jcho710	jcho710	@JoeBiden COVID VACCINE https://t.co/rSkDyDF1ia
16588	jcho710	jcho710	@JoeBiden COVID VACCINE https://t.co/rSkDyDF1ia
16662	jcho710	jcho710	@PressSec COVID VACCINE https://t.co/rSkDyDF1ia

As the web urls have been removed in our data cleansing part prior to emotion detection, the algorithm detects the text as neutral. That is the reason why the emotion color of the edges is Orange in the network.

Emotion Network Visualisation Using Customised Twitter API Data

In this tweet-centric small-scale dataset collected on May 3 and 11 from the high-level view, the three major nodes with high degree are found.

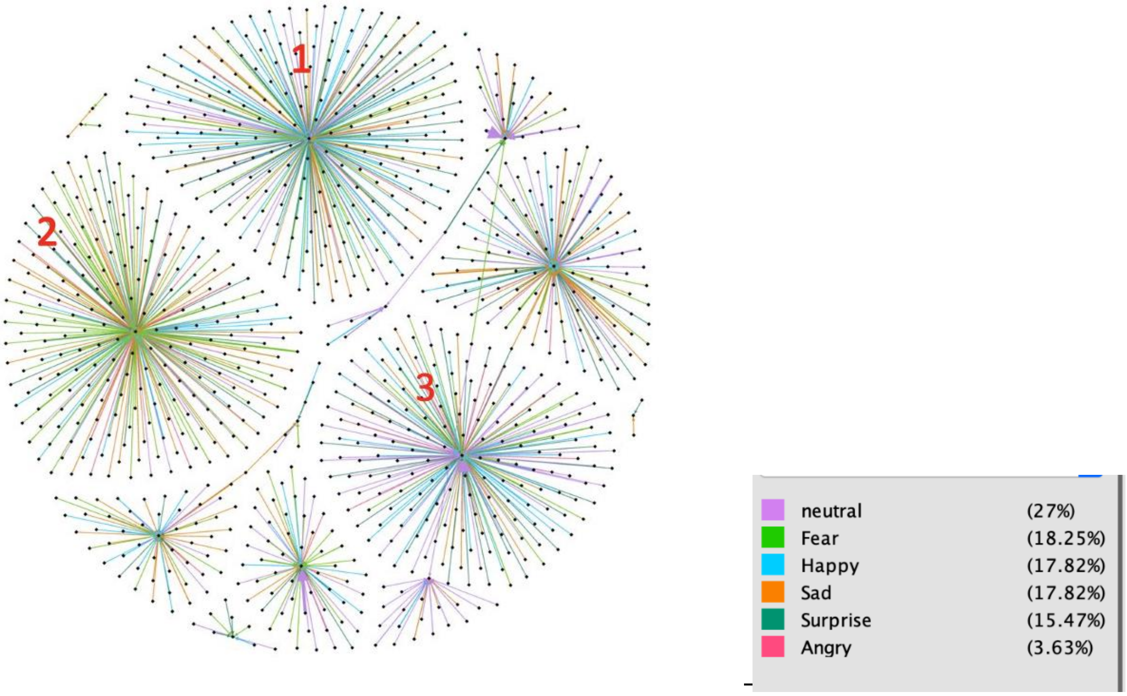
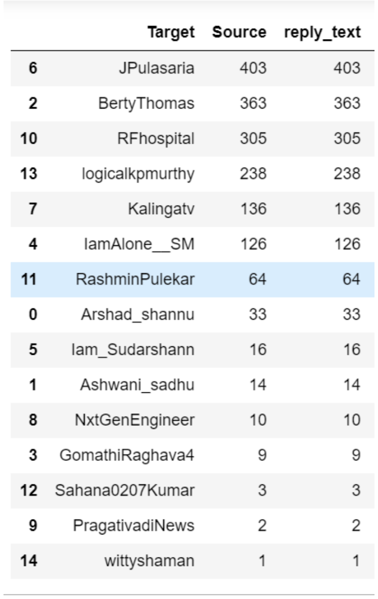


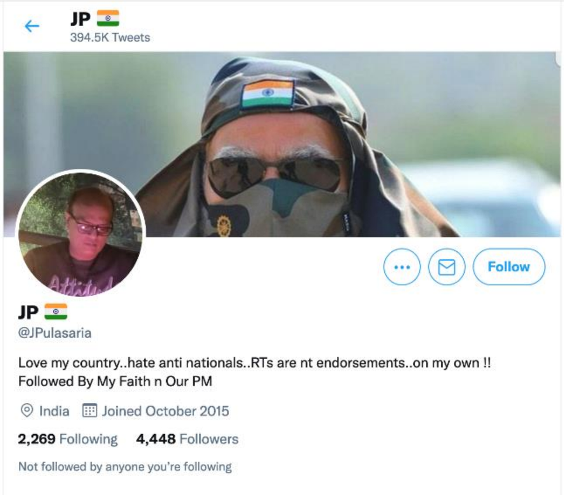
Figure 21. May 2020 Tweets emotion network visualisation

When we examined the edge count to understand the degree of node, the result in descending order is shown as below. The user @JPulaasria, @BertyThomas and @RFhospital are top three users with highest degree.



	Target	Source	reply_text
6	JPulasaria	403	403
2	BertyThomas	363	363
10	RFhospital	305	305
13	logicalkpmurthy	238	238
7	Kalingatv	136	136
4	IamAlone__SM	126	126
11	RashminPulekar	64	64
0	Arshad_shannu	33	33
5	Iam_Sudarshann	16	16
1	Ashwani_sadhu	14	14
8	NxtGenEngineer	10	10
3	GomathiRaghava4	9	9
12	Sahana0207Kumar	3	3
9	PragativadiNews	2	2
14	wittyshaman	1	1

In Twitter, the account @JPulaasria is found as middle-age strong believer in Prime Minister of India, Narendara Modi. In our dataset he expressed his optimism in the mid of COVID second wave in the country by posting "***Hoping to get 18 to 44 age group vaccination slots before I turn 45 #COVIDVaccination***". And he received 403 replies within the same day at the time of scraping. It is labelled as "1" in red in the network.The bluish emotion color (Represents Happy of his tweet's replies in the visualization which illustrates that his optimism was spread in his community and the positive vibe was maintained regardless of 4205 deaths in India on the date posted 11 May 2021.



JP 🇮🇳
394.5K Tweets

JP 🇮🇳
@JPulasaria

Love my country..hate anti nationals..RTs are nt endorsements..on my own !!
Followed By My Faith n Our PM

India Joined October 2015

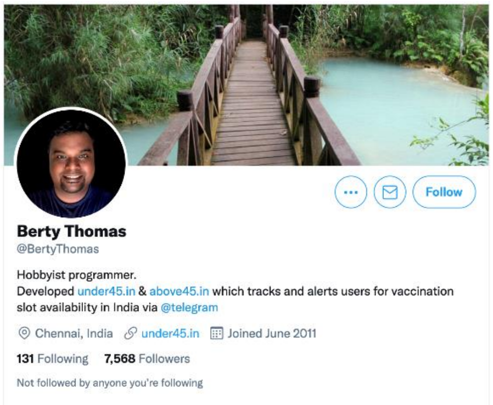
2,269 Following 4,448 Followers

Not followed by anyone you're following

Another interesting thing we found out is that in the second giant node by @bertyThomas, the tweet-replied text reflects the public experiencing the second wave of covid in India and expressing the cries for hospital beds and shortage of oxygen cylinder and rushing to register for vaccination online.

The tweet made by a user @BertyThomas received 363 replies (edges) within the same day. This surpasses the replies received by the hospital called @RFhospital. The replies he received are mixed with fear (green) and surprise(surprise). In the dataset, he posted "*Came for my vaccination at #Apollo (Greams road) for the 9-11am slot. The entire day's crowd is here. Token system*." His replies expressed the frustrations for instance "*Slots getting full within second*." The node of his post in the network can be seen surrounded by mostly green colour edges representing Fear. It is labelled as "2" in red in the network.

In twitter, this user @BertyThomas is found as just a programmer and not a public figure. He has been actively sharing the vaccination alerts and news in the community time to time every day and people are seriously following his posts to rush to get the vaccination slot. The replies expressed desperation for vaccine and fear of not getting the vaccine registration as the slots got full within a few seconds.



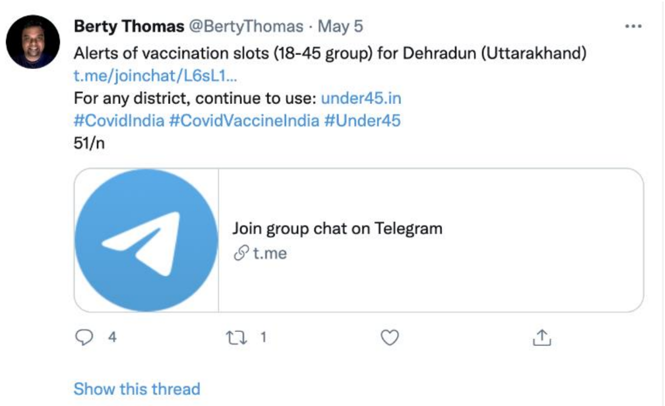
Berty Thomas
@BertyThomas

Hobbyist programmer.
Developed under45.in & above45.in which tracks and alerts users for vaccination slot availability in India via @telegram

Chennai, India under45.in Joined June 2011

131 Following **7,568** Followers

Not followed by anyone you're following



Berty Thomas @BertyThomas · May 5
Alerts of vaccination slots (18-45 group) for Dehradun (Uttarakhand)
t.me/joinchat/L6sL1...
For any district, continue to use: under45.in
#CovidIndia #CovidVaccineIndia #Under45
51/n

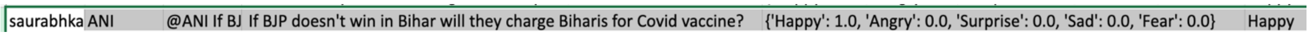
Join group chat on Telegram
t.me

Limitation of the Emotion Analysis

Researchers have identified different dimension of emotions from different perspectives. As the human emotions are complexed, it still has challenges to detect the correct emotions (Sailunaz & Alhajj, 2019). A single piece of text can present multiple emotions of a human, there is no perfect algorithm to identify the exact emotions.

In our analysis, the emotion results are produced by the library. However, there is no accuracy measures were conducted and we were not able to quantify the accuracy and errors of the emotion detection results. From the manual scanning the scores, most of the interpretation were correct and some texts were wrongly scored as shown in the screen shot below.

The user@ saurabhkadam_07 posted “*If BJP doesn't win in Bihar will they charge Biharis for Covid vaccine?”* The text expressed the political concern as well as the concern for vaccination in one of the states in India, Bihar. BJP in the text refers to the currently ruling party, Bharatiya Janata Party in India. The text2emotion library interpreted as “Happy” emotion.



saurabhka	ANI	@ANI If BJ	If BJP doesn't win in Bihar will they charge Biharis for Covid vaccine?	{'Happy': 1.0, 'Angry': 0.0, 'Surprise': 0.0, 'Sad': 0.0, 'Fear': 0.0}	Happy

Learning and Insights

From the emotion network analysis, we learned that sentiment can be spread easily in the community and a particular emotion could evolve based on interactions even on social media. It also worth noting how complex the conversations are during this period of a pandemic – as we see sentiments ranging not just on Fear or Sadness, but also Surprise, Anger, and Happiness.

References

Aguilar-Gallegos, N. (2020). *Dataset on dynamics of Coronavirus on Twitter*. https://doi.org/10.17632/7ph4nx8hnc.1

Cucinotta, D., & Vanelli, M. (2020). WHO Declares COVID-19 a Pandemic. *Acta Biomed, 91*, 157-160. https://doi.org/10.23750/abm.v91i1.9397

GIRVAN, M. & NEWMAN, M. E. J. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences,* 99**,** 7821-7826.

Karron, R. (2020). *A Top Vaccine Expert Answers Important Questions About a COVID-19 Vaccine*. John Hopkins School of Public Health. Retrieved 2021-05-18 from https://www.jhsph.edu/covid-19/articles/a-top-vaccine-expert-answers-important-questions-about-a-covid-19-vaccine.html

Khurshudyan, I., & Johnson, C. Y. (2020). *Russia unveils coronavirus vaccine 'Sputnik V,' claiming breakthrough in global race before final testing complete*. Washington Post. Retrieved 2021-05-18 from https://www.washingtonpost.com/world/russia-unveils-coronavirus-vaccine-claiming-victory-in-global-race-before-final-testing-is-complete/2020/08/11/792f8a54-d813-11ea-a788-2ce86ce81129_story.html

Sailunaz, K., & Alhajj, R. (2019). Emotion and sentiment analysis from Twitter text. *Journal of Computational Science, 36*, 101003. https://doi.org/10.1016/j.jocs.2019.05.009

Twitter. (2021). *About verified accounts*. Twitter. Retrieved 2021-05-18 from https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts

Wicke, P., & Bolognesi, M. M. (2021). Covid-19 Discourse on Twitter: How the Topics, Sentiments, Subjectivity, and Figurative Frames Changed Over Time. *Frontiers in Communication, 6*. https://doi.org/10.3389/fcomm.2021.651997

Appendix

Covid Vaccine tweets data dictionary (kaggle)

Feature	Description
Id	Unique identifier of tweets
conversation_id	Unique identifier of conversation within a tweet
created_at	Datetime of creation of tweets
date	Date of creation of tweets
time	Time of creation of tweets
timezone	Timezone info for feature “date” and “time”
user_id	Unique identifier of twitter account
username	Username of twitter account
name	Displayname of twitter account
place	Location of twitter account
tweet	Content of tweets
language	Language of tweets
mentions	List of twitter accounts mentioned in the tweet
urls	Website link or URL showed in the tweet
photos	Link or URL of photo showed in the tweet
replies_count	Number of replies of tweets
retweets_count	Number of retweets of tweets
likes_count	Number of likes of tweets
hashtags	List of hashtags
cashtags	List of cashtags
link	Link of a tweet
retweet	Retweet details
quote_url	Quoted url inside a tweet
video	Value to indicate tweets contain video or not, 1 for yes, 0 for no
thumbnail	URL of thumbnail
near	Names of location record where tweets are post
geo	Geolocations record where tweets are post
source	Unknown feature
user_rt_id	Unknown feature
user_rt	Unknown feature
retweet_id	Twitter IDs retweeted the tweet
reply_to	Twitter IDs replied the tweet
retweet_date	Date of retweet
translate	Unknown feature
trans_src	Unknown feature
trans_dest	Unknown feature

Dropbox link for code used: (will be sent via email as well)
https://www.dropbox.com/home/Covid19%20tweets%20code-SINA