Technical University of Ilmenau
Institute of Media and Communication Science, Fakultät für Informatik und Automatisierung
Data Science in Social Media: Graph Analysis and Text Mining in Practice
Jun.-Prof. Dr. Emese Domahidi, Prof.Dr Kai-Uwe Sattler, Aliya Andrich, Dr. Nadine Steinmetz

# Who are the opinion leaders on COVID-19? :

*A social graph formation and sentiment analysis*

Grau Chopite, Jessica (62359)
Tanaji Abhang, Komal(62529)
Yazdipour, Shahriar (62366)

# 1. Introduction

A respiratory disease, given the name of Coronavirus 19 (COVID-19), that causes acute respiratory issues, has spread worldwide and has caused multiple drastic changes in the world (Fauci et al., 2020). Important information has been transmitted both through mass media and online media. In fact, online platforms become even more prominent tools to understand the social discussion about the virus (Ferrara, 2020).

Some particular users from these social networks have been able to successfully convey and spread more messages across social networks. These users could be called opinion leaders. Traditionally, opinion leaders are a minority of members within a society that possess certain ideal qualities to persuade and spread ideas to others. However, with the widespread use of social media, even ordinary users can produce and effectively spread information (Meeyoung Cha et al., 2010).

Moreover, Twitter has played an important role within users' connections during crisis situations and in fact, COVID-19 has been widely discussed in this social network (Chen et al., 2020). Opinion leaders help change social norms and accelerate behavior change (Valente & Pumpuang, 2007).

## 1.1 Relevance

According to Worldometer, by September 24, 2020, more than 32 million cases of COVID-19 have been registered since the beginning of the pandemic and from that number, almost a million people have died of complications from the disease (Worldometer, 2020). Although the survival rate is quite high, all over the world measures to avoid the spreading of the virus have been taken. However, several months after the pandemic, there are still many skeptical people.

Governments and authorities have had to implement decisive and harsh laws and obligations in order to tackle the pandemic. These changes, however, have provoked strong skepticism, fear of being misled, and overall conspiracy theories of all sorts (Freeman et al., 2020). Information flow can influence others' attitudes (at least those expressed in the form of tweets), and said influence can be measured on Twitter (Xu et al., 2014). Therefore opinion leaders are good elements of strategies for the adoption of healthy measures (Locock et al.,

2001). It is important to know who and how the opinion leaders are in order to carry out important messages of global interests such as this pandemic.
Therefore the platform represents an opportunity to explore the attitude and expressions regarding the novel COVID-19. Furthermore, it offers a chance to see who the opinion leaders regarding this topic are and possible explanations on why they are precisely "heard".

### 1.2 Structure

In this project, a literature review was made in order to explore background information relevant for the work; later, the methods and results are thoroughly explained and finally, a discussion and outlook for future research are offered.

## 2. Literature review and research question

In the following chapter, a literature review on opinion leaders on Twitter, health communications on Twitter, and the relevance of Twitter during this health crisis originated by Coronavirus are further explained in order to give an overview of the topics in question.

### 2.1 Literature review

#### 2.1.1 Opinion leaders in Twitter

Information is spread throughout a social network by the persuasion from opinion leaders (Atkins et al., 2008). As mentioned in the previous sub-chapter, Twitter is a social network that can effectively diffuse information. Most of the time, opinion leadership can be seen in this platform and these users are usually the ones with high connectivity (Xu et al., 2014). In fact, the potential of Twitter is such that Park (2013), in a study about Twitter and its influence on the involvement in politics, stated that opinion leadership makes a difference in the involvement of users in complex topics such as politics (Park, 2013).

#### 2.1.2 Health-related communications on Twitter.

The impact of Twitter on public health is such that according to Paul & Dredze, (2011) in their paper *You Are What You Tweet: Analyzing Twitter for Public Health,* this social network alone has broad applicability for public health research. This is because not only do users share their opinions online but also symptoms or diseases they are currently going through which allows to more or less surveil possible signs of pandemics (Paul & Dredze,

2011). Furthermore, Twitter can also be a channel to promote behavior change regarding the intake or myths of medicines such as antibiotics (Scanfeld et al., 2010).
However, the effectiveness of Twitter regarding the spreading of health-related information can be a double-edged sword. This is the case of the popular anti-vaccine groups (anti-vax), whose arguments have been observed and efficiently raised in this social network (Gargiulo et al., 2020). This reaffirms the importance of understanding who and how the diffusion of information works on Twitter regarding the massively discussed Coronavirus.

### 2.1.3 COVID-19 on Twitter

The urgency and impact of the new Coronavirus have provoked the rapid development of research. In fact, Chen, Lerman & Ferrara (2020) were able to compile a dataset on Twitter which will hopefully facilitate the study of communication dynamics on COVID-19 (Chen et al., 2020). Although social media cannot replace the work of public health authorities, a study by Park, Park & Chong (2020) suggested that the sharing of news articles on medical information on COVID-19 was greater than the sharing of nonmedical news. The researchers also suggest that monitoring these public interactions and media news could help the urgently required decision making of health professionals (Park et al., 2020).
Throughout the literature review it is proved the importance of the information diffusion on COVID-19, and therefore, the importance of identifying the opinion leaders who could spread proper and helpful information regarding the virus. Moreover, this would help to defuse lies or reluctant behavior across the social network.

## 2.2 Research Question

After taking into consideration the background studies performed regarding the novel Coronavirus and the use of Twitter as an effective channel to explore the research on opinion leaders, the researchers of the current project aim to answer the following research question: Who are the opinion leaders on COVID-19?

# 3. Method

In this study, we propose a comprehensive method to detect opinion leaders of a COVID-19 topic on the Twitter platform. This study is based on the analysis of tweets related to COVID-19 which were tweeted in January, March, and April 2020. These tweets were not specific to COVID-19 tweets. Therefore, in the first step, COVID-19 tweets need to be

separated from irrelevant tweets. After filtering COVID-19 data, we extracted the relationship between Twitter users in terms of reply characteristic of Twitter - because reply relation includes interaction between users, which is the key factor of opinion leader detection - and generated a directed social graph. Social network graphs are analyzed according to centrality measures which show user's connectivity and involvement in the issue in order to find the opinion leader candidates during a pandemic period. The whole methodology is summarized in Figure 3.1.

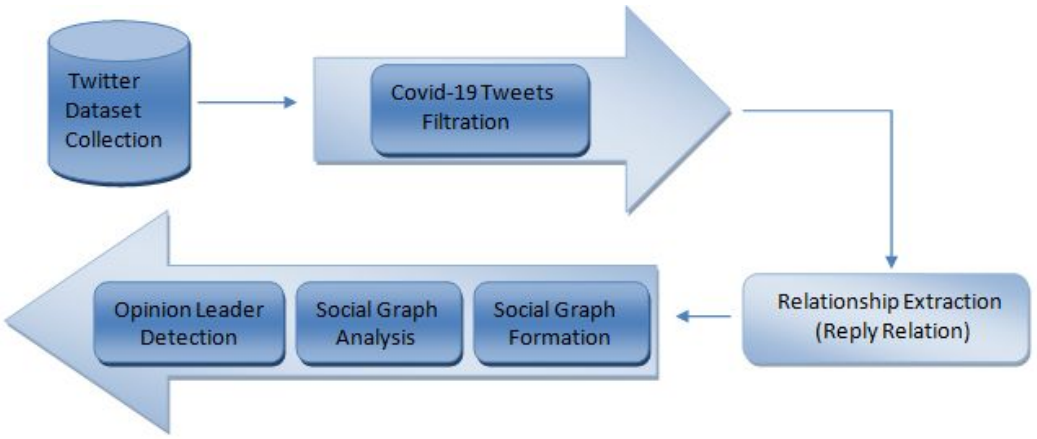


*Figure 3.1. Opinion Leader Detection Methodology*

### 3.1 Dataset Description and Filtering

The sample of this study is composed of tweets related to COVID-19. The Twitter dataset provides the information regarding the user's geolocation, time, and date of the tweet and also indicates if the tweet is a reply to another tweet. This includes COVID-19 tweets and other tweets that sum up to 207 Gigabytes of data. Initially, the dataset is divided into 12 different files, therefore, with the help of the multiprocessing concept, we used 12 different processes and create new threads for each 10000 chunk of records to filter the COVID-19 tweets based on tweet content, using Regex filters and drop irrelevant features to reduce data size. By saving the result of each filtering process thread into separate files we were able to use Dask Library to apply parallelization in our post processing steps. This method helped us to use the full capacity of our systems to reduce data filtering time significantly. At the end we were able to reduce the dataset to 400 Megabytes.

### 3.2 Approach to identify Opinion Leaders

The proposed approach consists of four steps. The first one is relationship extraction from tweets which refer to the determination of interactions between users. The second step is the social graph formation in which directed graphs are generated. The third step comes with the social network analysis part in which centrality metrics are applied to the social graphs and the final step is opinion leader detection.

#### 3.2.1 Relationship Extraction

In our dataset, we have the tweet text with reply-to-status-id and reply-to-user- id objects. Tweets are text-based messages and kind of summaries of opinions and ideas, therefore, those tweets were extracted and analyzed to detect the opinion leaders. A tweet may be an individual tweet about our topic or a reply to some other tweet. If the representative tweet is an individual tweet then reply-to-status-id and user-id objects will contain null. However, if the representative tweet is a reply then reply-to-status-id and reply-to-user-id will contain the original tweet's id and author id respectively. Since the social graph consists of users as nodes, the relationships between nodes should represent the interactions between these users. With the help of the given reply objects in the dataset, we can extract the following user reply relation:

Source Node: User id of the user who replied to the original tweet

Destination Node: Original user id (Reply-to-user-id)

#### 3.2.2 Social Graph Formation

By using the reply relationships as mentioned above, we created a relationship graph whose nodes are Twitter users. While forming a graph about a specific topic like COVID-19, a user is a node in the social graph if they post a tweet about that topic and have an interaction with any other node in the graph. The tweet may be an original tweet or a reply about COVID-19. If user A replies to user B and the reply tweet of A is related to our topic then there is an edge between nodes A and B in the social graph.

This social network graph formed between Twitter users is directional and the direction is determined as follows: If user A replies to user B: A → B

### 3.2.3 Social Network Analysis

After social graph formation, the social network analysis part comes in order to determine the opinion leaders of COVID-19 in the Twitter community. There are different social network analysis techniques available in the literature (Landherr et al, 2010).. In our research, we used centrality measurements, which denote the importance of a node in a graph, to analyze the generated social network graph. We used centrality measures such as degree centrality, betweenness centrality, and closeness centrality to find out the most influential node in the social network graph. In the following, we first formally define the above three centrality metrics one at a time and illustrate them by means of an example network.

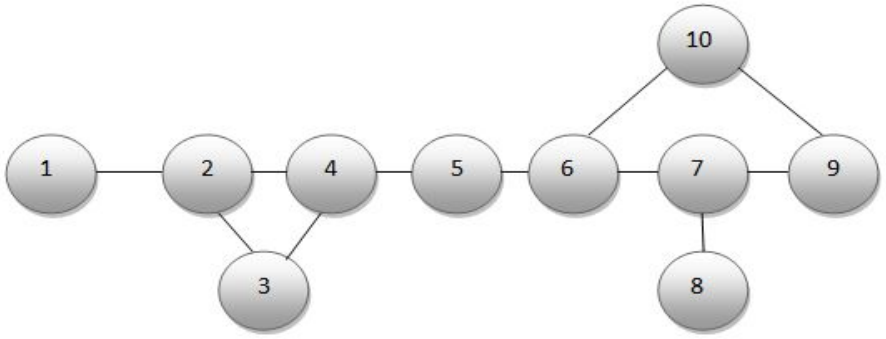


*Figure 3.2 Example of network for the illustration of centrality measures*

**Degree Centrality**

Degree centrality represents the simplest centrality metric and defines the number of direct contacts as an indicator of the quality of a user's interconnectedness in the network (Landherr et al, 2010). In our study, if any particular user gets the highest degree centrality score, it means that user got the highest number of replies from another user. In the network of Fig. 3.2, it follows that DC(1) = 1, since user 1 has only one direct relationship with user 2. However, user 4 has the centrality score of DC(4) = 3. Table 1 shows the values of DC for all nodes of the example network. In addition, the user's ranking is stated, i.e. their order in descending value of DC. The nodes 2, 4, 6, and 7 take rank 1 and, thus, are the best-networked nodes when applying this degree centrality measure.

**Table 1**

*Results for degree centrality*

| Nodes | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| DC(Node) | 1 | 3 | 2 | 3 | 2 | 3 | 3 | 1 | 2 | 2 |
| Rank | 3 | 1 | 2 | 1 | 2 | 1 | 1 | 3 | 2 | 2 |

**Betweenness Centrality**

In the case of betweenness centrality, a network node is considered to be well connected if that node is located on as many of the shortest paths as possible between pairs of other nodes. The underlying assumption of this centrality metric is that the interaction between two no directly connected nodes A and B depends on the nodes between A and B. In our study, the highest betweenness centrality of any particular user represents the strategic location of that node in the network. In the network of Fig.3.2, the betweenness centrality of user 9 is 1 since user 9 is located on one of the two shortest paths from user 7 and 8 to user 10. The values of the betweenness centrality for the other nodes and their ranking are listed in Table 2.

**Table 2**

*Results for Betweenness centrality*

| Nodes | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| BC(Node) | 0 | 8 | 0 | 18 | 20 | 21 | 11 | 0 | 1 | 6 |
| Rank | 8 | 5 | 8 | 3 | 2 | 1 | 4 | 8 | 7 | 6 |

**Closeness Centrality**

Closeness centrality is based on the idea that nodes with a short distance to other nodes can disseminate information very productively through the network (Landherr et al, 2010). In order to calculate the closeness centrality of a node A, the distances between the node A and all other nodes of the network are summed up (Landherr et al, 2010). By using reciprocal value we show that closeness centrality value increases when the distance to another node is reduced, i.e. when the integration into the network is improved. In our study, if any particular user gets the highest closeness centrality, it means that the user spreads information very productively across the network. For user 4 in the network of Fig. 3.2, results of closeness centrality is 1/21. This is because of the fact that for the users x = 2, 3, 5, the distance(4,x) =

1, moreover for the users x = 1, 6, the distance(4, x) = 2, for the users x = 7, 10, the distance(4, x) = 3 and for the users x = 8, 9, the distance(4, x) = 4 holds. Table 3 includes the closeness centrality scores of all nodes in the network in Fig. 3 and their ranking when applying closeness centrality.

**Table 3**

*Results for Closeness Centrality*

| Nodes | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| CC(Node) | 1/34 | 1/26 | 1/27 | 1/21 | 1/19 | 1/19 | 1/23 | 1/31 | 1/29 | 1/25 |
| Rank | 10 | 6 | 7 | 3 | 1 | 1 | 4 | 9 | 8 | 5 |

### 3.2.4 Opinion Leader Detection

We applied the degree, betweenness, and closeness centrality metrics to analyze the graph generated from the Twitter social network and detect the opinion leaders of Covid-19. After applying the three types of centrality measurements to the Twitter social graph, the resulting values of centrality measures are sorted and ranked. Degree centrality is measured as the degree of the given node and reflects the popularity and relational activity of a node and betweenness algorithms represents the node's ability to influence or control interaction between nodes it links in the network, therefore the combination of them may result in better and more accurate opinion leader identification (Santos et al., 2006). Moreover, closeness centrality measurements are based on the geodesic distance d(A,B) that is, the minimum length of the path from node A to node B. It indicates the user's availability, safety, and security (Frank, 2002). Therefore, we considered these measures as essential to identify the most influential users on COVID-19 from Twitter.

## 4. Results

In this study, first we filtered COVID-19 tweets from the received twitter dataset between January, March and April 2020. After filtering covid-19 data, we pre-processed the filtered dataset and extracted 17 different features from it.

After preprocessing steps, we extracted the reply relationship between Twitter users using tweet_in_reply_to_user_id as target node and user_id as source node. Even after filtering the COVID-19 dataset, we got a huge dataset which contains 2,80,949 nodes and 2,36,856 edges, for which social graph plotting is very difficult. Therefore we identified a list of users who got replies more than threshold value (200) using some mathematical calculations (i.e. basic averaging and counting functions were performed). These threw the results shown in Appendix B.

**Table 5**

*Top 5 OLs in terms of degree centrality and betweenness centrality*

| User ID | User name | Rank | Description | DC* (user_id) | BC* (user_id) |
|---|---|---|---|---|---|
| 25073877 | realDonaldTrump | 1 | President of the United States | 0.368531 | 0.660924 |
| 216299334 | piersmorgan | 2 | British journalist and TV host | 0.058994 | 0.105062 |
| 170699708 | weijia | 3 | Chinese American Journalist for CBS | 0.048838 | 0.082808 |
| 3131144855 | BorisJohnson | 4 | Prime Minister of the UK | 0.046299 | 0.080116 |
| 759251 | CNN | 5 | American News Channel | 0.043237 | 0.066690 |

*Note.* DC* stands for Degree Centrality and BC* stands for Betweenness Centrality.

Later, after getting these results, we made a list of users who got the highest replies and calculated the centrality measures for that users only. After calculating three different centrality measures for that user list, we sorted centrality metrics values and got the top five users as shown in Tables 5 and 6. We got exactly the same results for degree and betweenness centrality as they have similarities in their measurement styles and working concept and slightly different results for closeness centrality.

**Table 6**

*Top 5 OLs in terms of Closeness centrality*

| User ID | User name | Rank | Description | CC* (user_id) |
|---|---|---|---|---|
| 25073877 | realDonaldTrump | 1 | President of the United States | 0.44244 |
| 216299334 | tomvell | 2 | Chef and Owner of Spice n Flavor (indian restaurant) in California, USA | 0.35378 |
| 170699708 | LiesMistruths | 3 | -- | 0.34935 |
| 3131144855 | Hjoey15 | 4 | -- | 0.34308 |
| 759251 | grayjonv | 5 | -- | 0.34291 |

*Note.* CC* stands for Closeness centrality. The blank spaces marked with -- indicate that users' identities are not verified.

It can be seen that @realDonaldTrump is the user that remains on top on the three centrality measures with 0.368531 for DC, 0.660924 for BC and 0.44244 for CC. However, when it comes to the CC, CNN, Boris Johnson, Weijia Yang and Piers Morgan, these public figures' tweets regarding COVID-19 were not the fastest to spread, as they were not in the Top 5 opinion leaders. In contrast, not so well known figures like Thomas Vellaringattu (@tomvell) and ordinary users like the so-called Whistleblower (@LiesMistruths), @Hjoey15 and Jony gray (@grayjony) managed to get their tweets regarding COVID-19 spread faster.

## 5. Discussion and outlook

During the studied period in which the pandemics outbreak was at its peak, information regarding COVID-19 was widely discussed on Twitter. The detection of opinion leaders throughout this paper was possible thanks to a social graph, social network analysis and the centrality measures which gave us more insight regarding their identity. As a main result, it is possible to see that a big part of the opinion leaders were actually popular public figures. Donald Trump, whose account is @realDonaldTrump on Twitter, was found to have the highest centrality measures (degree centrality, closeness centrality and betweenness centrality) which indicates that his tweets regarding COVID-19 were the fastest to spread, were the most visible ones and had the most power; and were the ones that worked as a "bridge" between other users.

However, even though other prominent characters like Boris Johnson (@BorisJohnson), Piers Morgan (@piersmorgan), Weijia Jang (@weijia) or the news organization CNN (@CNN) had high degree centrality and betweenness centrality, they were apparently not the ones the fastest to spread (closeness centrality). They did, however, work as a connection for other users and they did have much influence over the network.
Therefore, it could be said that according to the results of this paper, the most prominent figures were indeed the opinion leaders, but not all of their content was as quick to spread as other ordinary users' tweets.

### 5.1 Outlook for future research

For further research, it would be interesting to perform a sentiment analysis in order to understand which sentiments were expressed by the opinion leaders to effectively spread their content regarding COVID-19. This would give place to the possibility of further studying a way on how to spread information regarding such an important global crisis quickly and efficiently.
Another further consideration for future work is to compare centrality measures and perhaps including other ones such as Eigenvector within this comparison in order to identify Opinion leaders more accurately.

### 5.2 Limitations

One of the main limitations for this research was the access to the data. For instance, Researchers do not have access to the reply count of each tweet and to get those data an enterprise license is required. Lack of properties like the number of retweets and quotes in the dataset also limited the possible methods that could have been used in this research project.
Another found limitation is that certain users, that were available in the Twitter data, no longer possessed an active profile on Twitter.
Moreover, it was found that when inspecting the tweet characteristics of said profiles, their amount of followers would change. For instance, in the information of one tweet it would be shown that the amount of followers was 2 and later on, was 10. Although this could be a result of the chronological difference between tweets and a consequence of their impact on other users, this could be resulting in a misinterpretation of the data.
Lastly, a limitation of this paper was the absence of a full comparison of centrality measures, as said in the subchapter 5.1, it would be a good addition for further research.

## References

Atkins, M. S., Frazier, S. L., Leathers, S. J., Graczyk, P. A., Talbott, E., Jakobsons, L., Adil, J. A., Marinez-Lora, A., Demirtas, H., Gibbons, R. B., & Bell, C. C. (2008). Teacher key opinion leaders and mental health consultation in low-income urban schools. *Journal of Consulting and Clinical Psychology*, *76*(5), 905–908. https://doi.org/10.1037/a0013036

Chen, E., Lerman, K., & Ferrara, E. (2020). Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health and Surveillance*, *6*(2), e19273. https://doi.org/10.2196/19273

Fauci, A. S., Lane, H. C., & Redfield, R. R. (2020). Covid-19 - Navigating the Uncharted. *The New England Journal of Medicine*, *382*(13), 1268–1269.

Ferrara, E. (2020). What Types of COVID-19 Conspiracies are Populated by Twitter Bots? *First Monday.* Advance online publication. https://doi.org/10.5210/fm.v25i6.10633

Frank, O. (2002). Using centrality modeling in network surveys. Social Networks, 24(4), 385–394. https://doi.org/10.1016/S0378-8733(02)00014-X

Freeman, D., Waite, F., Rosebrock, L., Petit, A., Causier, C., East, A., Jenner, L., Teale, A.‐L., Carr, L., Mulhall, S., Bold, E., & Lambe, S. (2020). Coronavirus conspiracy beliefs, mistrust, and compliance with government guidelines in England. *Psychological Medicine*, 1–13.

Gargiulo, F., Cafiero, F., Guille-Escuret, P., Seror, V., & Ward, J. K. (2020). Asymmetric participation of defenders and critics of vaccines to debates on French-speaking Twitter. *Scientific Reports*, *10*(1), 6599.

Landherr, A., Friedl, B. & Heidemann, J. A Critical Review of Centrality Measures in Social Networks. Bus Inf Syst Eng 2, 371–385 (2010). https://doi.org/10.1007/s12599-010-0127-3

Locock, L., Dopson, S., Chambers, D., & Gabbay, J. (2001). Understanding the role of opinion leaders in improving clinical effectiveness. *Social Science & Medicine*, *53*(6), 745–757. https://doi.org/10.1016/S0277-9536(00)00387-7

Meeyoung Cha, Hamed Haddadi, Fabrício Benevenuto, & Gummadi, K. P. (2010). Measuring user influence in twitter: The million follower fallacy. In *ICWSM 2010 - Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*.

Park, C. S. (2013). Does Twitter motivate involvement in politics? Tweeting, opinion leadership, and political engagement. *Computers in Human Behavior*, *29*(4), 1641–1648. https://doi.org/10.1016/j.chb.2013.01.044 (Computers in Human Behavior, 29(4), 1641-1648).

Park, H. W., Park, S., & Chong, M. (2020). Conversations and Medical News Frames on Twitter: Infodemiological Study on COVID-19 in South Korea. *Journal of Medical Internet Research*, *22*(5), e18897.

Paul, M.J., & Dredze, M. (2011). You Are What You Tweet:: Analyzing Twitter for Public Health. *International AAAI Conference on Web and Social Media*.

Santos, E. E., Pan, L., Arendt, D., & Pittkin, M. (2006). An Effective Anytime Anywhere Parallel Approach for Centrality Measurements in Social Network Analysis. In IEEE International Conference on Systems, Man and Cybernetics, 2006. SMC '06 (pp. 4693–4698). IEEE / Institute of Electrical and Electronics Engineers Incorporated. https://doi.org/10.1109/ICSMC.2006.385045

Scanfeld, D., Scanfeld, V., & Larson, E. L. (2010). Dissemination of health information through social networks: Twitter and antibiotics. *American Journal of Infection Control*, *38*(3), 182–188. https://doi.org/10.1016/j.ajic.2009.11.004

Valente, T. W., & Pumpuang, P. (2007). Identifying opinion leaders to promote behavior change. *Health Education & Behavior: The Official Publication of the Society for Public Health Education*, *34*(6), 881–896. https://doi.org/10.1177/1090198106297855

Worldometer. (2020). *COVID-19 Coronavirus Pandemic*. https://www.worldometers.info/coronavirus/

Xu, W. W., Sang, Y., Blasiola, S., & Park, H. W. (2014). Predicting Opinion Leaders in Twitter Activism Networks. *American Behavioral Scientist*, *58*(10), 1278–1293. https://doi.org/10.1177/0002764214527091

## Appendix A: Script file

See script file.

## Appendix B: Table 4. Results of data pre-processing and mathematical calculation

**Table 4**

*Results of data pre-processing and mathematical calculation*

| | |
|---|---|
| List of Extracted Features: | Tweet_id , tweet_text, tweet_tags, tweet_created_at, tweet_in_reply_status_id, tweet_in_reply_to_user_id, place_id, place_country, place_name, user_id, user_name, user_friends_count, user_fav_count, user_created_at, user_verified, user_statuses_count, user_followers_count |

| | |
|---|---|
| Number of users who get replies | 146099 |
| Top users (users with more than 200 replies) | 44 |
| Number of tweets the top users generated | 3092 |
| Number of replies to top users' tweets | 24216 |