



Studying leaders & their concerns using online social media during the times of crisis - A COVID case study

Rahul Goel¹ · Rajesh Sharma¹

Received: 1 April 2021 / Revised: 4 May 2021 / Accepted: 6 May 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Austria, part of Springer Nature 2021

Abstract

Online social media (OSM) has emerged as a prominent platform for debate on a wide range of issues. Even celebrities and public figures often share their opinions on a variety of topics through OSM platforms. One such subject that has gained a lot of coverage on Twitter is the Novel Coronavirus, officially known as COVID-19, which has become a pandemic and has sparked a crisis in human history. In this study, we examine 29 million tweets over three months to study highly influential users, whom we refer to as leaders. We recognize these leaders through social network techniques and analyse their tweets using text analysis. Using a community detection algorithm, we categorize these leaders into four clusters: *research*, *news*, *health*, and *politics*, with each cluster containing Twitter handles (accounts) of individual users or organizations. e.g., the *health* cluster includes the World Health Organization (@WHO), the Director-General of WHO (@DrTedros), and so on. The emotion analysis reveals that (i) all clusters show an equal amount of *fear* in their tweets, (ii) *research* and *news* clusters display more *sadness* than others, and (iii) *health* and *politics* clusters are attempting to win public *trust*. According to the text analysis, the (i) *research* cluster is more concerned with recognizing *symptoms* and the development of *vaccination*; (ii) *news* and *politics* clusters are mostly concerned with *travel*. We then show that we can use our findings to classify tweets into clusters with a score of 96% AUC ROC.

Keywords COVID-19 · Social Network · OSM · PageRank · Community detection · LDA · Sentiment analysis

1 Introduction

A leader is a person who can reach and inspire a large number of people with her/his opinions (Goleman 2004). Crisis times provide an opportunity to test these leaders and assess their ability to take timely and necessary actions (Littlefield and Quenette 2007; Schroeder et al. 2012; Bruns et al. 2013; Littlefield and Quenette 2007). Nowadays, leaders often use online social media (OSM) platforms to communicate information to crisis-affected people efficiently. During a crisis, such as the current COVID-19 pandemic, communication is critical since it influences people's opinions, attitudes, and psychological conditions. Many public figures worldwide have acknowledged that the infodemic of disinformation is a

significant secondary crisis caused by the pandemic, including UN Secretary-General (Pastor-Escuredo and Tarazona 2020). Infodemics can exacerbate the pandemic's real-world adverse effects across many dimensions, including social, physical, and even sanitary. Thus, leaders often ensure that accurate information is widely disseminated, and that misinformation is clarified using OSM.

This work centered on Twitter as it has attracted more than 628 million tweets related to COVID-19 by May 8, 2020¹ primarily because the COVID-19 pandemic affected more than 180 countries and territories worldwide (CSSE 2020). It impacted the lives of many individuals by keeping them in isolation or total lockdown. During these conditions, individuals are inclined to use social media platforms to share their experiences or concerns.

1.1 Study highlights

In this paper, we analyse 29 million tweets from 6 million unique users, collected through Twitter's API by using

✉ Rahul Goel
rahul.goel@ut.ee
Rajesh Sharma
rajesh.sharma@ut.ee

¹ Institute of Computer Science, University of Tartu, Tartu, Estonia

¹ <https://www.tweetbinder.com/blog/covid-19-coronavirus-twitter/>

COVID-19 trending keywords on Twitter (“coronavirus”, “coronavirusoutbreak”, and “COVID-19”). Our dataset spans from February 01, 2020, to May 02, 2020, and this study covers the following:

1. **Identifying leaders:** We recognize the leaders during the COVID-19 crisis by creating users’ tweet network and using social network analysis techniques. We define *Leaders* as users with high PageRank values. Please note that *leaders* can either be an individual user or an organization (e.g., the *World Health Organization*). However, we consider them as a separate entity and do not distinguish between them. We use Girvan-Newman community detection method to identify the optimal number of communities. Based on the highest modularity value, we categorize leaders into four clusters²: *research*, *news*, *health*, and *politics* (refer Section 4 for detail). We further study these clusters by analyzing their tweets to reveal their inclination and concerns during COVID-19.
2. **Leader’s concerns:** Our finding shows that leaders are mostly discussing about five topics that are related to (1) *symptoms*; (2) *vaccination*; (3) *hygiene*; (4) *travel*; and (5) *pandemic* (Section 4). We also use various text analysis techniques to understand clusters’ alignment toward these concerns during the COVID-19 crisis. We observe that *research* cluster is more concerned about understanding the *symptoms* and development of *vaccination*; *news* and *politics* clusters about *travel* and *hygiene*. Finally, using Chi square independent test, we show that the number of concerns (or topics) and the type of leading cluster are dependent (Section 5).
3. **Classification model:** Based on our findings, we build a multi-class predictive model for estimating the probability that a tweet belongs to a specific leading cluster. Our model can classify tweets among leading clusters with a score of 96% AUC ROC (Section 6).

1.2 Contributions

The contribution of our analysis is three-fold:

1. Previous COVID-19 studies of leadership used social media to propose a framework to characterize leaders based on nodes centrality (Pastor-Escuredo and Tarazona 2020; Rufai and Bunce 2020). In contrast, this paper identifies leaders’ and their clustering using community detection algorithms and the PageRank centrality method.

2. Next, we show that the number of public concerns (or topics), and the type of leading cluster is dependent on one another using Chi-square independent test with a p-value of $1.411e-33$.
3. Finally, while previous research has only identified leaders, we use a machine learning approach to classify tweets into relevant clusters using tweets’ features such as tweet’s text, its sentiment, and topic.

1.3 Paper organization

The rest of the paper is as follows. Next, we discuss the related work. We then describe the dataset in Sect. 3. Section 4 presents the study of leaders during COVID-19, and their alignments toward various concerns are covered in Sect. 5. In Sect. 6, we use various machine learning models to classify tweets into various leading clusters with good accuracy. We conclude with a discussion of future directions in Sect. 7.

2 Related work

Our work lies at the intersection of leadership, social media, and the COVID-19 pandemic. Thus, in this section, we present literature concerning various works which have studied leadership from different viewpoints. Broadly, studies on leadership types during COVID-19 can be divided into the following: (1) Religious leaders, (2) Academic leaders, and (3) Information leaders.

2.1 Religious leaders during Covid-19

Religious leaders of faith-based communities played a crucial role in supporting individuals, families, and communities in coping with the pandemic, especially those who were ill or bereaved by COVID-19 (Greene et al. 2020). To acknowledge that COVID-19 is a global pandemic, affecting all races, ethnicities, and geographic regions, various global religious and inter-religious groups have issued guidance, advisory notes, and statements to support the actions and roles of religious leaders during the pandemic (Organization 2020). In Yezli and Khan (2021), the authors believe that temporary closure of places of worship for group prayers and religious services should be implemented worldwide (especially in countries with local COVID-19 transmission) regardless of faiths involved.

In another work, the authors describe the religiously innovative adaptations made to customary rituals by Jewish religious leaders to fight COVID-19 (Frei-Landau 2020). These adaptations included allowing spiritual prayer through a “balcony”, conducting online prayer sessions using video conferencing, and broadcasting the Passover ceremony. In

² Note: In this work, “leader” represents a Twitter handle, and “cluster” represents a group of Twitter handles.

Dein et al. (2020), the authors examined anxiety and distress among members of a Modern Orthodox Jewish community to be quarantined in the USA. Their study demonstrates that public health authorities have an opportunity to form partnerships with religious leaders to promote health and psycho-social support, protect communities against stigma and discrimination.

Additionally, several works discussed the impact of virtual religious activities on mental health (Dein et al. 2020). The authors showed that limited empirical research is available on religion and mental health, and they have relied heavily on newspaper and internet sources. It has also been studied that healthcare professionals are often unprepared to answer the patients' religious beliefs regarding the diseases (Hashmi et al. 2020), which results in religious clichés and stigma among patients. Therefore, the inclusion and collaboration of spiritual leaders with healthcare professionals are needed to ensure a holistic understanding and overcome the stigma that can shape as a barrier for reaching an optimal therapeutic outcome (Hashmi et al. 2020).

2.2 Academic leaders during Covid-19

Academic leaders worldwide have responded by moving their educational and associated activities online; as a sense of immediacy swept the nation. The decision to pivot to remote learning was made swiftly, particularly by those institutions operating a shared leadership model (Fernandez and Shaw 2020). A survey study sought to understand the impact of the COVID-19 pandemic on school leadership in K-12 schools in a South Texas region indicate that school leaders were generally confident in their preparedness to best serve students, staff, and parents during the COVID-19 pandemic but felt a lack of resources and a preponderance of student inequities complicated the experience (Varela and Fedynich 2020).

In Brammer and Clark (2020), the authors discussed that COVID-19 has profound impacts on tertiary education globally. Border closures, cuts to aviation capacity, mandatory quarantine on entering a country, restrictions on mass gatherings, and social distancing all pose challenges to higher education institutions. However, there are possible solutions for Academic Leadership to improve diversity, equity, and inclusion (DEI) in the workplace, at community and institute levels, and in broader policy and decision-making (Maas et al. 2020).

2.3 Information leaders during pandemics

In the past, researchers have analysed Twitter data for studying various epidemic outbreaks such as Ebola (Oyeyemi

et al. 2014; Carter 2014), H1N1 (Smith 2010), flu (Achrekar et al. 2011), swine flu (Ritterman et al. 2009), etc. In Chew and Eysenbach (2010), the authors illustrate the potential of monitoring public health using social media during the H1N1 pandemic. In another work (Signorini et al. 2011), authors track the disease level and public concern during the H1N1 pandemic in the US using Twitter data.

A set of work has also focused on understanding the epidemic spreading patterns at different geographic locations using social media data and call record data (Goel and Sharma 2020; Goel et al. 2019). They also studied effect of COVID on stock market (Goel et al. 2020) using social media data. In Jain and Kumar (2015), authors track the spread of influenza-A (H1N1) in India using Twitter data. Vaccination and anti-viral uptake during H1N1 in the UK are studied in McNeill et al. (2016) using tweets. In Kumar and Kumar (2018), the authors examined the Nipah virus in India. In a very recent work (Samaras et al. 2020), researchers compared Google and Twitter data for predicting influenza cases in Greece.

Information is vital during a crisis such as the current COVID-19 pandemic as it significantly shapes people's opinion, behavior, and even their psychological state. The Secretary-General of the United Nations has acknowledged that the infodemic of misinformation is an essential secondary crisis produced by the pandemic. Infodemics can amplify the pandemic's real negative consequences in different dimensions: social, economic, and even sanitary (Pastor-Escuredo and Tarazona 2020). For instance, infodemics can lead to hatred between population groups that fragment the society, influencing its response or result in negative habits that help the pandemic propagate. On the contrary, reliable and trustful information along with messages of hope and solidarity can be used to control the pandemic, build safety nets and help promote resilience and anti-fragility. Several works proposed frameworks to characterize Twitter leaders based on the social graph analysis derived from the activity in this social network (Pastor-Escuredo and Tarazona 2020; Rufai and Bunce 2020; Lee et al. 2020; Goel and Sharma 2021).

In this work, we analyzed online social media, particularly Twitter, to identify leaders using centrality and community detection algorithms during COVID-19. Additionally, we grouped these leaders into four major clusters (health organizations, politics, news, and research) and analyzed each cluster's tweet to identify their respective concerns using natural language processing technique, specifically using Latent Dirichlet Allocation (LDA) method with Gibbs sampling. Finally, we created a classification model using a machine learning technique to cluster tweets.

3 Dataset description

This section first provides information about the coronavirus (officially known as COVID-19), and Twitter, an online social media platform. Next, the data collection and pre-processing are discussed in the following subsections.

3.1 COVID-19 and social media

In the last decade, humanity has faced many different pandemics such as SARS, H1N1, and presently COVID-19 caused by the SARS-CoV-2 coronavirus (Goel and Sharma 2020). The severity of these pandemics can be understood by the death toll claimed by them. The COVID-19 pandemic, which started in December 2019 from Wuhan (Organization 2020; Huang et al. 2020), China has infected 128,253,282 individuals and claimed 2,804,677 (as of March 30th, 2021) deaths worldwide (CSSE 2020). This impacted individuals life by keeping them in isolation, or total lockdown. During these lockdown conditions, individuals are inclined to use social media platforms to share their thoughts and experiences. By May 8, 2020, Twitter has attracted more than 628 million tweets related to coronavirus which makes it a significant online social media platform for COVID-19 study³.

3.2 Twitter

Twitter is an online social media platform classified as a microblogging site with which users can share messages, links to external Web sites, images, or videos that are visible to their followers. Messages that are posted are short in contrast to traditional blogs. Blogging becomes ‘micro’ by shrinking it down to its bare essence and relaying the heart of the message and communicating the necessary as quickly as possible in real-time. Twitter, in 2016, limited its messages to 140 characters (Giachanou and Crestani 2016) and presently the maximum length of its message (also called a tweet) has been increased to 280 characters⁴. Apart from Twitter, there are other microblogging platforms such as Tumblr⁵, FourSquare⁶, Google+⁷, and LinkedIn⁸ of which Twitter is the most popular microblogging site. In the past, researches have shown that Twitter data is well suited for sentiment analysis and opinion mining (Tumasjan et al. 2010).

³ <https://www.tweetbinder.com/blog/covid-19-coronavirus-twitter/>

⁴ <https://developer.twitter.com/en/docs/basics/counting-characters>

⁵ <https://www.tumblr.com/>

⁶ <https://foursquare.com/>

⁷ <http://plus.google.com>

⁸ <http://linkedin.com/>

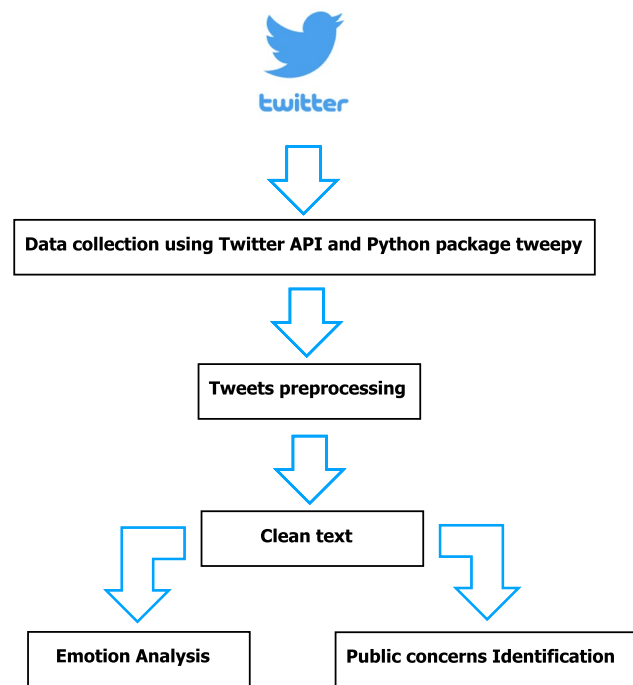


Fig. 1 Tweets collection, preprocessing, emotions and public concern identification flow

3.3 Data collection from twitter

The data collection is done using Twitter Streaming Application Programming Interface (API) and Python (see Fig. 1). API is a tool that facilitates the interaction between computer programs and Web services. It enables the real-time collection of data by tracking the live stream of public tweets. Many Web services provide developers with APIs for interacting with their services and for programmatically accessing data. Python library such as *tweepy*⁹ also assist this task by providing functions that can track live streams of public tweets using hashtags or usernames.

In this work, we use the Twitter Streaming API and Python package *tweepy* to download tweets related to various keywords regarding rapidly increasing coronavirus disease (COVID-19). We collect and store a large sample of public tweets beginning February 1, 2020, that matched a set of pre-defined keywords: “coronavirus” and “coronavirusoutbreak”. The additional keyword is later added such as “COVID-19” (On February 11, 2020, COVID-19 was the official name given to novel coronavirus by WHO). The objective of using these specific keywords is to collect tweets belonging to COVID-19. The Twitter data are stored in JSON format to make it easy for parsing and analysis.

⁹ <https://www.tweepy.org/>

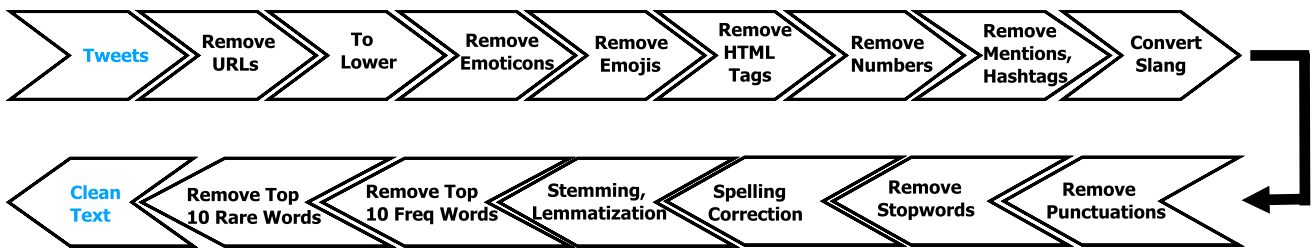


Fig. 2 Tweets preprocessing

Table 1 Description of the dataset

Parameter	Value
Time period	01-02-2020 to 02-05-2020
#Tweets	29,469,349
#Original tweets	6,494,657
#Retweets	22,974,692
#Users	7,875,334
#Features	91

3.4 Tweets preprocessing

Figure 2 shows the basic steps that we took in preprocessing the tweet dataset. We start our tweet preprocessing by removing all the URL's as they do not include any useful information for text analysis. Afterward, the tweets are subjected to lowercasing. The emojis, emoticon, numbers, mentions (@), and hashtags (#) are excluded from the tweets due to their specific semantics in tweets and overall text. In addition, all slang text are converted into its real meaning to understand the real context of the tweet.

Furthermore, we proceeded with removing every punctuation marks and a variety of different stopwords available in *nltk* package. We also perform the spell correction using *SpellChecker* package in Python. Using *nltk* package, we did the stemming and lemmatization of the tweet text. The dataset's size in terms of total number of tweets, original tweets, retweets, number of distinct users, and the number of distinct features are shown in Table 1. Twitter features include information about tweets and users. This encapsulates core attributes that describe each tweet data such as author, message, unique ID, a timestamp of when it was posted, and sometimes geo metadata shared by the user. Each user information like Twitter name, ID, number of followers, and most often account biography are also associated with tweets. With each tweet, we also get the tweet's common contents such as hashtags, mentions, media, and links.

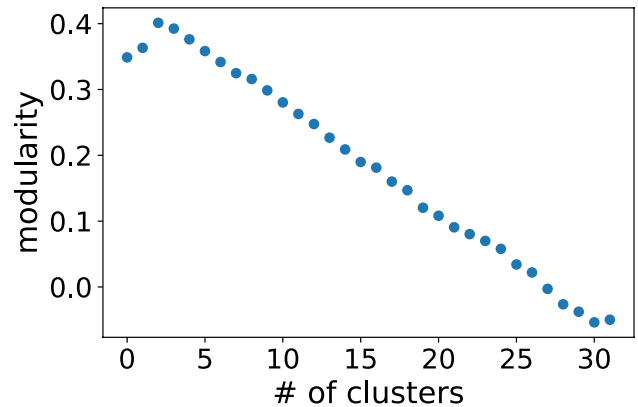


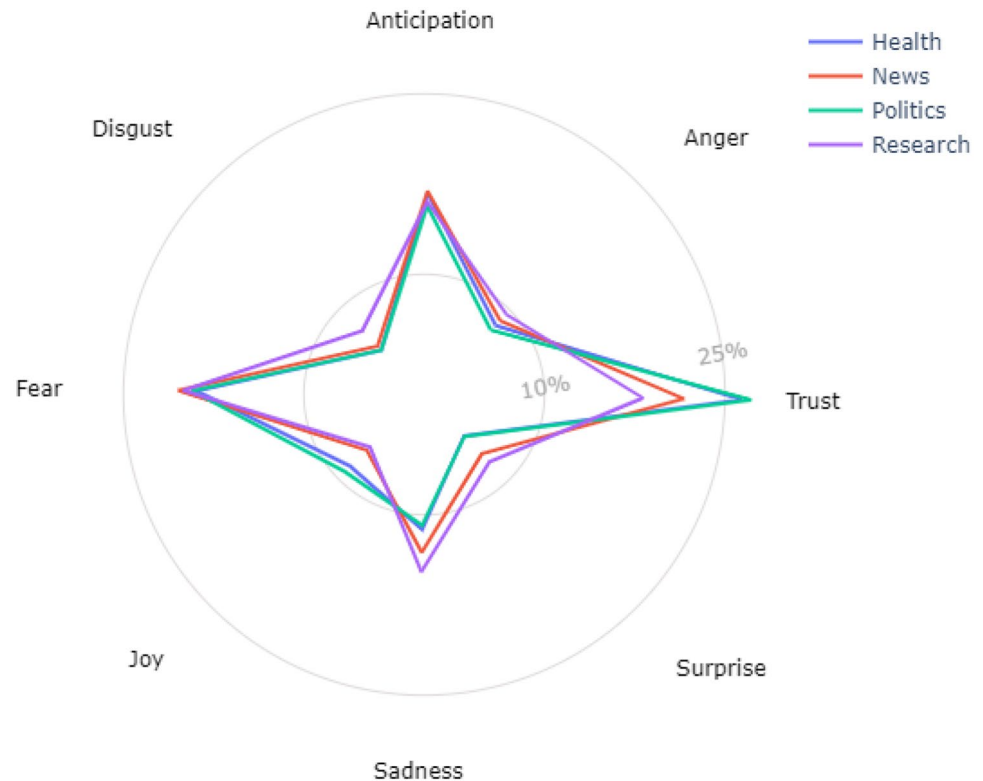
Fig. 3 Optimal number of clusters using Girvan-Newman method. Here, x-axis represents the number of clusters identified by Girvan-Newman community detection algorithm, and y-axis represents modularity value for different numbers of clusters. We find that the optimal value for the number of clusters is 4 (with highest modularity value) for our dataset

4 Studying leaders during COVID-19 crisis

To understand the users' interactions and to identify leaders, we first build a network by capturing the user's social connections. We build the directed "retweet" network among users where an edge $a \rightarrow b$ indicates that user a retweets user b . To identify users with similar alignment, we grouped them into communities using the Girvan-Newman method. Furthermore, to identify important nodes in the network, we employ the PageRank algorithm (Page et al. 1999), a well-known algorithm to characterize the centrality of nodes. PageRank reflects the importance of a node in the retweet network, and a higher PageRank value represents influential users who can spread their tweet content to a community much faster compared to users with lower PageRank value.

The number of optimal communities identified by Girvan-Newman method with highest modularity is four (see Fig. 3). With some manual inspection of influential users in each community, we categorized communities into four clusters, **Health**, **Politics**, **News** and **Research**, depending

Fig. 4 Dominant emotions in tweets from various leaders



upon their publicly available information such as profile description. Some of the examples of users under four different clusters are as follows:

1. *Health cluster*: which includes health related Twitter handlers such as *World Health Organization (@WHO)*, *Centers for Disease Control and Prevention (@CDCgov)*, *Director General WHO (@DrTedros)*, *Global Health Strategies (@GHS)* and *World Health Organization Western Pacific (@WHOWPRO)*.
2. *Politics*: includes Twitter profiles related to politicians such as *U.S. President (@realDonaldTrump)* and *U.S. House Candidate*.
3. *News*: examples of Twitter handlers include *China Global Television Network (@CGTNOfficial)*, *Al Jazeera English (@AJEnglish)* and *Global Times (@globaltimesnews)*.
4. *Research*: such as *ISCR*.

This indicate that, on Twitter, users are (re)sharing information from more reliable sources regarding COVID-19. Please note that some “leaders” are in fact teams, who tweet under a particular twitter handle, representing the official position of an office holder, e.g., Director General WHO(@DrTedros).

Dominant emotions for various clusters: Next, we analyse tweets to discover each clusters’ dominant emotions. As can be imagined, the majority of tweets from the Health,

Politics, News and Research clusters would be fact-based, our dataset includes individuals’ Twitter handles as well and these handles often express emotions that can be useful for our classification model. We use the *Syuzhet*¹⁰ package (Jockers 2015) available in R for emotions extractions. The *Syuzhet* package is designed to work on a sentence level and repeated words do not count toward the emotion assignment. We manually inspected a random sample of tweets for the accuracy of *Syuzhet* package and found it reasonably accurate with 83% accuracy. We used original tweets (without pre-processing) for the manual checking.

We observe that *Anticipation*, *Fear*, *Sadness*, and *Trust* are dominant emotions for all clusters compared to *Anger*, *Disgust*, *Joy*, and *Surprise* (see Fig. 4). Furthermore, it can be observe that all clusters show similar amount of *fear* and *anticipation* in their tweets. *Political* and *health* clusters indicate higher *trust* compared to others. On the other hand, *researchers* and *news* clusters display greater *sadness* toward the COVID-19 situation. Therefore, we can infer that leaders are displaying *fear*, *anticipation*, and *sadness* but at the same time, they are trying to build a *trust* environment among public.

Furthermore, we analyzed the leaders’ dominant emotions over time. This is shown using Fig. 5a (for February), 5b (for March), and 5c (for April). We observe that there

¹⁰ <https://github.com/mjockers/syuzhet>

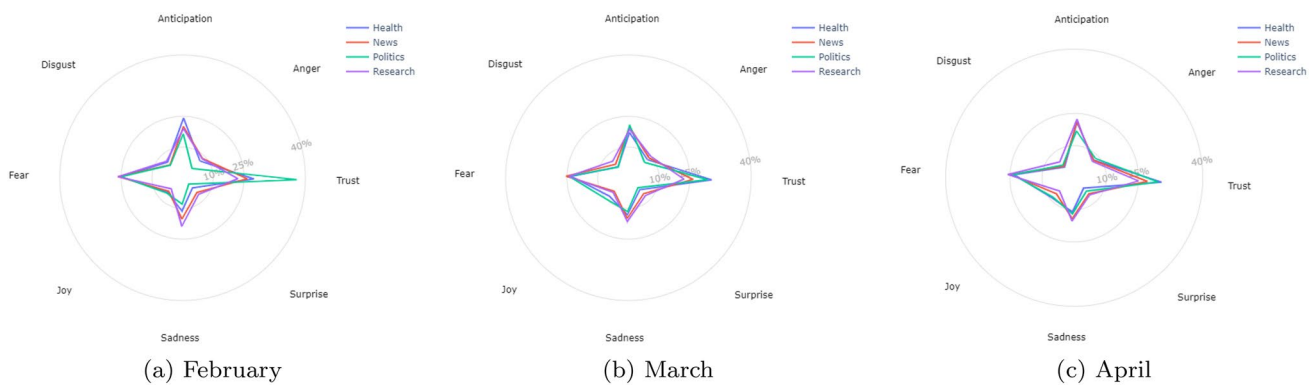


Fig. 5 Dominant emotions in tweets from various leaders over time

Table 2 Topic Modeling using LDA with Gibbs sampling

S.No.	Words	Topic (Proposed)
1	flu, symptomatic, fever, infected, spreading, china, cold, sick, severe, asymptomatic	Symptoms
2	china, effective, flu, cdc, world, research, testing, vaccine, pandemic, treatment	Vaccination
3	hand, wash, sanitizer, time, home, touch, masks, water, soap	Countermeasures
4	italy, international, quarantine, pandemic, traveling, restrictions, china, government, iran, world	Travel
5	china, stop, cdc, who, pandemic, public, death, rate, flu, news	Pandemic

is only minor changes in emotions of all clusters, except in the politics cluster. In February, the politics cluster were showing more *Trust*, but over time, it changed into *Sadness* and *Anticipation*. For other clusters, insignificant change in emotions is observed over time.

Topic modeling: In addition, we perform the text analysis to understand the focus of the interest of leaders during the COVID-19. We use the topic modeling technique to extract these interests. In particular, we use the Latent Dirichlet Allocation (LDA) Blei et al. (2003) method with Gibbs sampling (Porteous et al. 2008) which is an unsupervised and probabilistic machine-learning topic modeling method that extracts topics from text data. The key assumption behind LDA is that each given text is a mix of multiple topics.

We extracted between three to ten topics from tweets’ text and found that our dataset covers five different topics. To check this result and be sure on the number of topics, we compute two typical measurements used in LDA: perplexity and coherence scores for robustness check. The LDA returned a set of ten words related to each identified topic but not the title of the topic (see Table 2, Column 2). We then assign appropriate topics to each set of words that closely reflect the topic at an abstract level (Table 2, Column 3). It can be observed that leaders are mostly discussing topics related to various public concerns during the COVID-19. These concerns are: (1) *disease symptoms*, (2) *disease vaccination*, (3) *disease countermeasures or hygiene*, (4) *disease transmission during travel* and (5) COVID-19 as *pandemic/*

epidemic. In the next section, we analyse the clusters’ alignment toward these topics, which also reflects public concerns.

5 Clusters alignment toward various concerns

To study the alignment of various clusters toward five most discussed topics or concerns discovered in Sect. 4 regarding rapidly increasing coronavirus disease (COVID-19), we annotate each tweet using specific keywords. For example, disease symptoms tweets are annotated (i.e., using keyword *symptom*); vaccination (i.e., using keywords *vaccine* and *vaccination*); disease countermeasures (i.e., keywords *hygiene*, *wash*, *hand* and *mask*); disease transmission during travel (i.e., keywords *travel*, *flying*, *fly*, *airplane*, *flight* and *trip*) and pandemic (i.e., keywords *pandemic* and *epidemic*). In this section, we start by uncovering the specifics of these public concerns to understand them in more detail. We also explain the alignment of leading clusters toward these concerns.

5.1 Disease symptoms

Tweets analysis belonging to the symptoms category shows that the daily percentage of observed COVID-19 tweets related to symptoms is higher than other concerns, reflecting that users on Twitter are more concerned about the

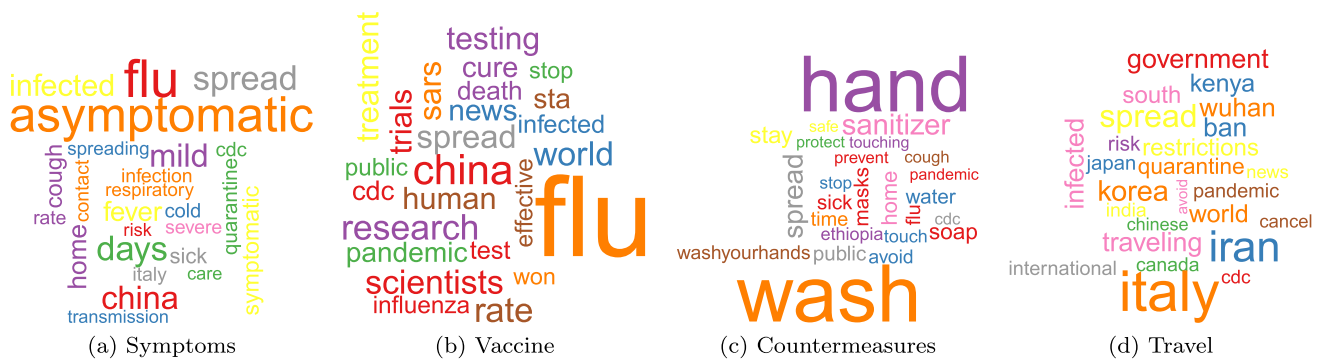


Fig. 6 Wordcloud

disease symptoms. Furthermore, we extract the coronavirus symptoms from tweets (see Fig. 6a). The commonly discussed symptoms are fever, cough, and cold. Leaders also discuss the similarity of COVID-19 and influenza/flu, as both spreads among people in a similar way, that is, via respiratory droplets from coughing or sneezing (Rothan and Byrareddy 2020). Leaders call the COVID-19 an asymptomatic disease as, in many cases, an infected person does not show any symptoms (Rothan and Byrareddy 2020). As anticipated, the *research* cluster is more concerned with understanding the COVID-19 symptoms than the other three clusters (see Fig. 7).

5.2 Disease vaccination

Our dataset spans from February 01, 2020 to May 02, 2020 and in that time period no vaccination or specific treatment for COVID-19 was available, therefore, leaders are discussing the effectiveness of flu vaccination for COVID-19 as both flu and COVID-19 cause respiratory disease (see Fig. 6b). We can also observe that leaders are discussing intensely the ongoing research to cure COVID-19. Therefore, we can infer that leaders on Twitter are very conscious of the symptoms of COVID-19 and also keeping an eye on the vaccination research for COVID-19. As no vaccine was available in that time period, all clusters are cautiously tweeting regarding the vaccination (see Fig. 7).

5.3 Disease countermeasures

For more detail analysis, we partition countermeasures tweets into two categories: (i) Hygiene (using keywords such as *hygiene*, *wash*, *hand*) and (ii) Mask (using keyword *mask*). The explanation behind this categorization is that these two countermeasures (*hygiene* and *mask*) have distinct essence. We find that the volume of tweets pertaining to *Mask* is higher than *Hygiene* before February 24, and after that leader started focusing more on *Hygiene* compared to *Mask*. This indicates that initially, leaders were tweeting

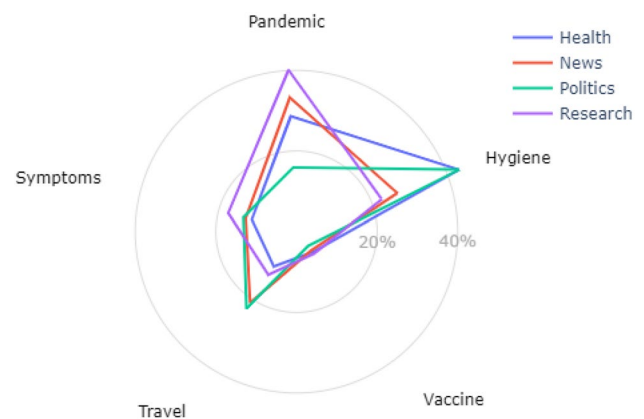


Fig. 7 Leaders alignment toward various public concerns

about wearing masks but later they shifted their countermeasures strategy toward taking proper *Hygiene* against COVID-19.

To explore the countermeasures against COVID-19, we created a wordcloud from countermeasures category tweets (see Fig. 6c) indicating suggestions such as handwashing, respiratory hygiene, self-isolation and self-quarantine as prevention. Hand wash using soap and water or sanitizer is highly discussed for countermeasures. Handwashing is also recommended by the US Centers for Disease Control and Prevention (CDC) to prevent the spread of the disease. It recommends that people should wash hands more often with soap and water for at least 20 seconds, especially after going to the toilet or when hands are visibly dirty, before eating, and after blowing one's nose, coughing, or sneezing. It further recommended using an alcohol-based hand sanitizer with at least 60% alcohol by volume when soap and water are not readily available¹¹. The advice from CDC

¹¹ <https://www.cdc.gov/coronavirus/2019-ncov/about/steps-when-sick.html>

to avoid touching the eyes, nose, or mouth with unwashed hands and for *respiratory hygiene*, cover the mouth with masks and take precautions in case of coughing¹² are also highly tweeted.

A high number of tweets belonging to *Health* and *politics* clusters are focused on hygiene compared to tweets from *news* and *research* clusters (see Fig. 7).

5.4 Disease transmission during traveling

To explore the effect of COVID-19 on traveling, we create a word cloud from travel category tweets (see Fig. 6d). This indicates that countries quarantined travelers from many countries to control pandemics or avoid spreading infection. Some of the mention countries either quarantined travelers from other countries or banned by others, such as China, Italy, South Korea, Japan, Iran, Canada, India, and Kenya. CDC has also issued guidelines for flight travelers and crew¹³. Among clusters, *news* and *politics* clusters are more focusing on traveling compared to *research* and *health* clusters (see Fig. 7). Although, it is interesting to observe that *politics* cluster is discussing more about *travel* compared to *news*.

5.5 COVID-19 as a pandemic

As COVID-19 outbreak is first identified in Wuhan, China, in December 2019 (Organization 2020). The World Health Organization (WHO) declared the outbreak a *Public Health Emergency of International Concern* on January 30th 2020 and a pandemic on March 11th 2020 (Organization 2020). Our analysis on tweets shows that *research*, *news* and *health* clusters are frequently using pandemic or epidemic keyword while discussing about COVID-19 compared to *politics* cluster (see Fig. 7). Interestingly, this reflects the non-supportive behavior of *politics* cluster toward considering COVID-19 as pandemic (Rothwell and Makridis 2020).

Next, we test the correlation of leading clusters with public concerns. Considering that both leading clusters and public concerns are categorical variables, we perform Chi square independent test. We begin with the null hypothesis (H_0) and alternate hypothesis (H_a) as:

- H_0 : The number of tweets related to public concerns is independent of the type of leading cluster.
- H_a : The number of tweets related to public concerns is dependent of the type of leading cluster.

¹² <https://www.cdc.gov/coronavirus/2019-ncov/travelers/index.html>

¹³ <https://www.cdc.gov/quarantine/air/managing-sick-travelers/ncov-airlines.html>

The Chi square independent test results a p-value of $1.411e-33$. Considering $\alpha = 0.05$, we observe that p-value is smaller than α . Since, $\alpha > p$ -value, we reject H_0 . This means that the factors are not independent. Hence, we can conclude that at a 5% level of significance, from the data, there is sufficient evidence to conclude that the number of tweets related to public concerns and the type of leading cluster are dependent on one another.

To summarize, on analyzing the clusters' alignment toward various public concerns (see Fig. 7), we find that *researchers* are highly concerned about understanding the COVID-19 by studying its symptoms and development of the vaccination. *News* are discussing travel and hygiene. *Health* cluster is focusing on hygiene. Whereas, *political* people are highly concerned about travel and hygiene. This indicates that the different clusters are focused on specific public concerns. This can be viewed as a positive approach since various clusters are focusing on particular issues and engaging with each other on common problems.

6 Tweet classification in clusters

Next, taking insights from previous sections, we build a model to estimate the likelihood that a tweet belongs to a specific cluster.

6.1 Features used for learning

To illustrate the predictive power of various feature sets, we define a series of models, each with a different feature set, as mentioned in the earlier section. We focus on three different types of features. One of them is "Tweet text" is part of the original dataset features, and others ("Emotions" and "public concerns") are extracted using tweet text. These features are the following:

1. **Tweet text:** This refers to the clean text extracted after preprocessing the original tweets (see Section 3.4).
2. **Emotions:** This refers to the emotions associated with respect to each tweet (see Section 4).
3. **Public concerns:** This corresponds to the public concerns revealed in Sect. 5.

6.2 Experimental setup and results

We aim to estimate the likelihood that a tweet belongs to a specific cluster using the features mentioned in Sect. 6.1. As in Sect. 4, we filter the leaders and cluster them into four groups. We remove all user tweets that are not in any of the leader clusters. We also deleted all blank tweets after tweet preprocessing. After applying these filters, we obtain a dataset containing 42,468 tweets.

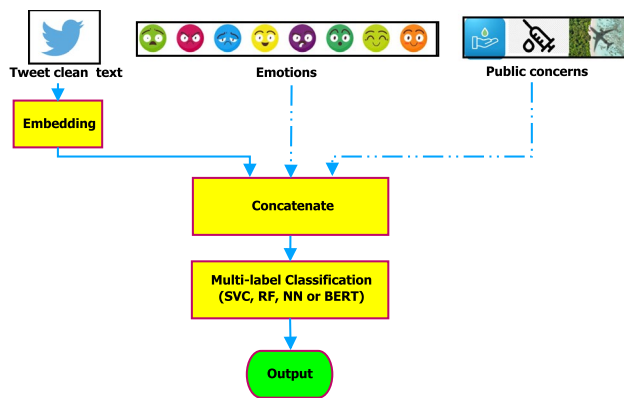


Fig. 8 Flow diagram for features concatenation and model selection

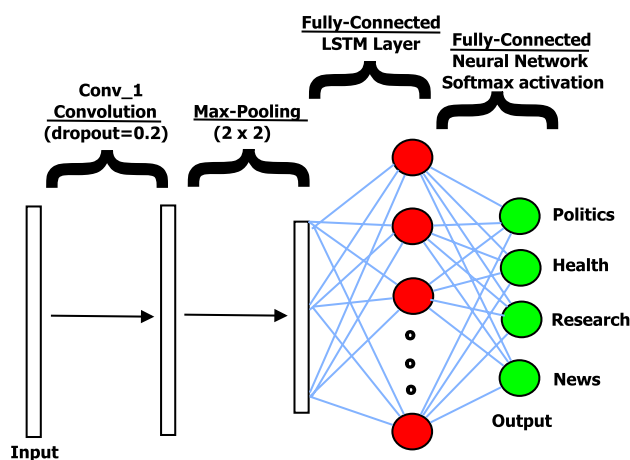


Fig. 9 Random neural network framework

The tweets percentage and count belonging to different clusters are as follows: *News* (36%; 15,289), *Health* (33%; 14,023), *Research* (18%; 7,635), and *Politics* (13%; 5,521). As tweet distribution among clusters is imbalanced, this problem is represented as an *imbalanced multi-class classification* (Zhang et al. 2015). Since the dataset is imbalanced, we use Synthetic Minority Over-sampling Technique (SMOTE) Chawla et al. (2002) to resolve this problem. SMOTE works by selecting examples close in the feature space, drawing a line between the examples in the feature space, and drawing a new sample at a point along that line. Specifically, a random example from the minority class is first chosen, then k of the nearest neighbors for that example is found (typically $k=5$). A randomly selected neighbor is chosen, and a synthetic example is created at a randomly selected point between the two examples in feature space.

We experiment with several classification models, including Support Vector Classifier (SVC) (Tsuda 1999; Lee and Lee 2007), Random Forests (RF) (Breiman 2001), Random

Table 3 Model mean AUC ROC and standard deviation for various dataset features. For example, the BERT model with the dataset (Text+Emotion) achieves 93.6% mean AUC ROC accuracy with 0.8% standard deviation

Data	SVC	RF	NN	BERT
Text	91.9 (2)	95.3 (2)	91.8 (9)	92.5 (8)
Text+Concerns	92.4 (3)	95.3 (2)	92.7 (9)	92.8 (9)
Text+Emotions	93.2 (3)	96.0 (2)	93.5 (9)	93.6 (8)
Text+Emotions + Concerns	93.6 (2)	96.0 (2)	94.1 (9)	95.2 (8)

neural network (NN, see Fig. 9 for framework) (Hansen and Salamon 1990; Alom et al. 2020), and Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2018). Figure 8 displays the data flow to the models. We find that the Random Forest model is the most effective for our task. As the dataset is imbalanced and the trade-off between true and false positive rates associated with classification, we choose to compare models using the area under the receiver operating characteristic (ROC) curve (AUC) (Davis and Goadrich 2006; Flach 2004). Thus, a random baseline will score 50% on AUC ROC. We use 5-fold cross-validation for estimation, and we run all models three times. Therefore, 15 iterations in total for each model. We also standardize *emotion* and *public concern* features to have zero mean and unit variance.

Results: Table 3 shows the classification accuracy for our models and the best result is highlighted in bold for each model. With the RF model trained on all available features, we achieve a mean accuracy of 96% AUC ROC with 0.2% as standard deviation. The low value of standard deviation indicates that the model is robust. All model shows that the most crucial feature for the classification is clean text. We can note that our best model, RF, is trained with the TF-IDF embedding for the words a user used in their tweets. We also see that public concerns and emotions are essential characteristics for classifying a tweet in different clusters.

7 Discussion and conclusion

With a quest to understand the role of various leaders during COVID-19, we study a large number of tweets using techniques such as network analysis, text analysis, and sentiment analysis. Based on network analysis, users are categorized into four different clusters: *research*, *news*, *health*, and *politics*. The results using text analysis shows that leaders of different clusters are focused on various public concerns. In particular, researchers about understanding the pandemic

symptoms and development of *vaccination*; news about *travel* and *hygiene*; health cluster about *hygiene*; and political individuals about *travel* and *hygiene*. Additionally, sentiment analysis indicated that emotions such as *anticipation*, *fear*, *sadness*, and *trust* dominate all clusters. Also, the evolution of emotions with time shows that there is only minor changes in emotions of all clusters, except in the politics cluster. In February, the politics cluster were showing more *Trust*, but over time, it changed into *Sadness* and *Anticipation*. Lastly, we showed that the extracted features could be used to identify tweets' clusters with an AUC ROC score of up to 96%.

Limitations. This work has some limitations, primarily related to generalization and users' response analysis on tweets. First, the analysis and classification model may not be generalized for the data collected in similar epidemic/pandemic situations (such as *#Ebola*). Our analysis is based upon users' posts on Twitter (i.e., tweets). Ideally, it would help us analyse the comments (that is, reply and retweets) regarding original tweets to get a broader understanding of the topic. Comments can be helpful to understand users' reactions to tweets. However, we showed that even an existing LDA model could effectively extract different topics from tweet text. Another limitation of our work is that the Twitter stream is filtered following Twitter's API documentation; hence the tweets analyzed here still constitute a representative subset of the stream instead of the entire stream.

Future work could consider several important directions to determine the effect of users' response and media (such as images and videos) in such pandemic situations. We intend to analyse the tweets' data to a greater extent for different category users over time to understand their change in tweet pattern and focused concerns. Future work should also explore more sophisticated models and deeply analyse various cluster writing styles to capture their tweets' signature.

Author Contributions *Rahul Goel*: Data curation, Formal analysis, Investigation, Validation, Visualization, and Writing - original draft. *Rajesh Sharma*: Supervision, and Writing - Review & Editing.

Funding This research is funded by ERDF via the IT Academy Research Programme, and H2020 framework project, SoBigData++, and CHIST-ERA project SAI.

Availability of data and material Our dataset is Twitter data, and it is not allowed to share this dataset publicly according to Twitter data sharing rules.

Declarations

Conflict interest The authors declare that they have no conflict of interest.

Code availability Not applicable. As this work utilizes a non-shareable dataset.

References

- Achrekar H, Gandhe A, Lazarus R, Yu SH, Liu B (2011) Predicting flu trends using twitter data. In: 2011 IEEE conference on computer communications workshops (INFOCOM WKSHPs). IEEE, pp 702–707
- Alom Z, Carminati B, Ferrari E (2020) A deep learning model for twitter spam detection. *Online Soc Netw Med* 18:100079
- Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *J Mach Learn Res* 3(Jan), 993–1022
- Brammer S, Clark T (2020) Covid-19 and management education: Reflections on challenges, opportunities, and potential futures. *Bri J Manag* 31(3):453
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
- Bruns A, Highfield T, Burgess J (2013) The arab spring and social media audiences: English and arabic twitter users and their networks. *Am Behav Sci* 57(7):871–898
- Carter, M (2014) How twitter may have helped nigeria contain ebola. *BMJ: Br Med J (Online)* 349
- Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority over-sampling technique. *J Artif Intell Res* 16:321–357
- Chew C, Eysenbach G (2010) Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak. *PloS one* 5(11)
- COVID C, Global cases by Johns Hopkins Center for Systems Science and Engineering (CSSE). <https://coronavirus.jhu.edu/map.html>
- CSSE J (2020) Coronavirus covid-19 global cases by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)
- Davis J, Goadrich M (2006) The relationship between precision-recall and roc curves. In: Proceedings of the 23rd international conference on Machine learning, pp 233–240
- Dein S, Loewenthal K, Lewis CA, Pargament KI (2020) Covid-19, mental health and religion: An agenda for future research
- Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) (2018)
- Fernandez AA, Shaw GP (2020) Academic leadership in a time of crisis: The coronavirus and covid-19. *J Leadersh Stud* 14(1):39–45
- Flach P (2004) The many faces of roc analysis in machine learning. ICML Tutorial
- Frei-Landau R (2020) when the going gets tough, the tough get-creative: Israeli jewish religious leaders find religiously innovative ways to preserve community members' sense of belonging and resilience during the covid-19 pandemic. *Psychological Trauma: Theory, Research, Practice, and Policy* 12(S1):S258
- Giachanou A, Crestani F (2016) Like it or not: A survey of twitter sentiment analysis methods. *ACM Comput Surv (CSUR)* 49(2):1–41
- Goel, R., Ford, L.J., Obrizan, M., Sharma, R.: Covid-19 and the stock market: evidence from twitter. arXiv preprint [arXiv:2011.0717](https://arxiv.org/abs/2011.0717) (2020)
- Goel, R., Sharma, R.(2020) Mobility based sir model for pandemics - with case study of covid-19. In: 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 110–117 . <https://doi.org/10.1109/ASONAM49781.2020.9381457>
- Goel, R., Sharma, R. (2021) Studying leaders during times of crisis using online social media—a covid case study. arXiv preprint [arXiv:2101.03002](https://arxiv.org/abs/2101.03002)
- Goel R, Singh A, Ghanbarnejad F (2019) Modeling competitive marketing strategies in social networks. *Physica A: Statis Mech Appl* 518:50–70
- Goleman D (2004) What makes a leader? *Harvard Bus Rev* 82(1):82–91

- Greene T, Bloomfield MA, Billings J (2020) Psychological trauma and moral injury in religious leaders during covid-19. *Theory, Research, Practice, and Policy, Psychological Trauma*
- Hansen LK, Salamon P (1990) Neural network ensembles. *IEEE Trans on pattern Anal Mach Intell* 12(10):993–1001
- Hashmi FK, Iqbal Q, Haque N, Saleem F (2020) Religious cliché and stigma: A brief response to overlooked barriers in covid-19 management. *J Religion Health* 59(6):2697–2700
- Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X et al (2020) Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The lancet* 395(10223):497–506
- Jain VK, Kumar S (2015) An effective approach to track levels of influenza-a (h1n1) pandemic in india using twitter. *Procedia Computer Science* 70:801–807
- Jockers, ML (2015) Syuzhet Extract sentiment and plot arcs from text. Retrieved October 21, 2015
- Kumar AA, Kumar AA (2018) Deadly nipah outbreak in kerala: Lessons learned for the future. *Indian journal of critical care medicine: peer-reviewed, official publication of Indian Society of Critical Care Medicine* 22(7):475
- Lee D, Lee J (2007) Domain described support vector classifier for multi-classification problems. *Pattern Recognition* 40(1):41–51
- Lee S, Hwang C, Moon MJ (2020) Policy learning and crisis policy-making: quadruple-loop learning and covid-19 responses in south korea. *Policy Soc* 39(3):363–381
- Littlefield RS, Quenette AM (2007) Crisis leadership and hurricane katrina: The portrayal of authority by the media in natural disasters. *J Appl Commun Res* 35(1):26–47
- Maas B, Grogan KE, Chirango Y, Harris N, Liévano-Latorre LF, McGuire KL, Moore AC, Ocampo-Ariza C, Palta MM, Perfecto I et al (2020) Academic leaders must support inclusive scientific communities during covid-19. *Nat Ecol Evol* 4(8):997–998
- McNeill A, Harris PR, Briggs P (2016) Twitter influence on uk vaccination and antiviral uptake during the 2009 h1n1 pandemic. *Frontiers in public health* 4:26
- Organization, WH (2020) Novel coronavirus—china
- Organization, WH (2020) Statement on the second meeting of the international health regulations (2005) emergency committee regarding the outbreak of novel coronavirus (2019-ncov)
- Organization WH et al (2020) Practical considerations and recommendations for religious leaders and faith-based communities in the context of covid-19: interim guidance, 7 April 2020. *World Health Organization, Tech. rep*
- Oyeyemi SO, Gabarron E, Wynn R (2014) Ebola, twitter, and misinformation: a dangerous combination? *Bmj* 349:g6178
- Page L, Brin S, Motwani R, Winograd T (1999) The pagerank citation ranking: Bringing order to the web. *Tech. rep, Stanford InfoLab*
- Pastor-Escuredo D, Tarazona, C (2020) Characterizing information leaders in twitter during covid-19 crisis. *arXiv preprint arXiv:2005.07266*
- Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P, Welling, M (2008) Fast collapsed gibbs sampling for latent dirichlet allocation. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 569–577
- Ritterman, J., Osborne, M., Klein, E.(2009) Using prediction markets and twitter to predict a swine flu pandemic. In: *1st international workshop on mining social media*, vol. 9, pp. 9–17
- Rothan, HA., Byrareddy, SN (2020) The epidemiology and pathogenesis of coronavirus disease (covid-19) outbreak. *J Autoimmunity* p. 102433
- Rothwell, J, Makridis, C (2020) Politics is wrecking america’s pandemic response. *Brookings Institution*. September 17, 2020
- Rufai SR, Bunce C (2020) World leaders’ usage of twitter in response to the covid-19 pandemic: a content analysis. *Journal of Public Health* 42(3):510–516
- Samaras L, García-Barriocanal E, Sicilia MA (2020) comparing social media and google to detect and predict severe epidemics. *Sci Rep* 10(1):1–11
- Schroeder R, Everton S, Shepherd R (2012) Mining twitter data from the arab spring. *Combat Terrorism Exch* 2(4):54–64
- Signorini A, Segre AM, Polgreen PM (2011) The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS one* 6(5)
- Smith, TO (2010) A little birdie told me: H1n1 information and misinformation exchange on twitter. *UALR-William H. Bowen School of Law Legal Studies Research Paper* (11-03)
- Tsuda K (1999) Support vector classifier with asymmetric kernel functions. In: *in European Symposium on Artificial Neural Networks (ESANN)*. Citeseer
- Tumasjan A, Sprenger TO, Sandner PG, Welpe IM (2010) Predicting elections with twitter: What 140 characters reveal about political sentiment. In: *Fourth international AAAI conference on weblogs and social media*
- Varela DG, Fedynich L (2020) Leading schools from a social distance: Surveying south texas school district leadership during the covid-19 pandemic. In: *National Forum of Educational Administration and Supervision Journal*, vol. 38
- Yezli S, Khan A (2021) Covid-19 pandemic: it is time to temporarily close places of worship and to suspend religious gatherings. *J Travel Med* 28(2), taaa065
- Zhang ML, Li YK, Liu XY (2015) Towards class-imbalance aware multi-label learning. In: *Twenty-Fourth International Joint Conference on Artificial Intelligence*