A text-mining analysis of public perceptions and topic modeling during the COVID-19 pandemic using Twitter data

Sakun Boon-Itt

Submitted to: Journal of Medical Internet Research
on: June 30, 2020

Table of Contents

A text-mining analysis of public perceptions and topic modeling during the COVID-19 pandemic using Twitter data

Sakun Boon-Itt[1] PhD

[1]Department of Operations Management, Center of Excellence in Operations and Information Management Thammasat Business School Thammasat University Bangkok TH

Corresponding Author:
Sakun Boon-Itt PhD
Department of Operations Management, Center of Excellence in Operations and Information Management
Thammasat Business School
Thammasat University
2 Prachan Road Pranakorn
Bangkok
TH

Abstract

Background: Coronavirus disease (COVID-19) is a scientifically and medically novel disease that is not fully understood as it needs to be consistently and deeply studied. In the past, research on the COVID-19 outbreak was only able to predict quantity data such as the number of outbreaks, but not infoveillance data.

Objective: This study aims to understand public perceptions on the trends of the COVID-19 pandemic and uncover meaningful themes of concern posted by Twitter users during the pandemic throughout the world.

Methods: Data mining on Twitter was conducted to collect a total of 107,990 tweets between December 13 and March 9, 2020. The analysis included time series, sentiment analysis and topic modeling to identify the most common topics in the tweets as well as to categorize clusters and find themes from keyword analysis.

Results: The results indicate three main aspects of public awareness and concerns regarding the COVID-19 pandemic. Firstly, the study indicated the trend of the spread and symptoms of COVID-19, which was divided into three stages. Secondly, the results of the sentiment analysis and emotional tendency showed that the people had a negative outlook toward COVID-19. Thirdly, topic modeling and themes relating to COVID-19 and the outbreak were divided into three categories, including (1) emergency of COVID-19 impact, (2) the epidemic situation and how to control it, and (3) news and social media reporting on the epidemic.

Conclusions: Sentiment analysis and topic modeling can produce useful information about the trend of COVID-19 pandemic and alternative perspectives to investigate the COVID-19 crisis which has created considerable public awareness around the world. This finding shows that Twitter is a good communication channel for understanding both public concern and awareness about COVID-19 disease. These findings can help health departments to communicate information as to what the public thinks about the disease.

(JMIR Preprints 30/06/2020:21978)
DOI: https://doi.org/10.2196/preprints.21978

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

A text-mining analysis of public perceptions and topic modeling during the COVID-19 pandemic using Twitter data

Sakun Boon-itt[1]*, PhD

1. Department of Operations Management, Center of Excellence in Operations and Information Management, Thammasat Business School, Thammasat University, Thailand
*Corresponding author: sboonitt@tu.ac.th

Abstract:

Background: Coronavirus disease (COVID-19) is a scientifically and medically novel disease that is not fully understood as it needs to be consistently and deeply studied. In the past, research on the COVID-19 outbreak was only able to predict quantity data such as the number of outbreaks, but not infoveillance data.

Objective: This study aims to understand public perceptions on the trends of the COVID-19 pandemic and uncover meaningful themes of concern posted by Twitter users during the pandemic throughout the world.

Methods: Data mining on Twitter was conducted to collect a total of 107,990 tweets between December 13 and March 9, 2020. The analysis included time series, sentiment analysis and topic modeling to identify the most common topics in the tweets as well as to categorize clusters and find themes from keyword analysis.

Results: The results indicate three main aspects of public awareness and concerns regarding the COVID-19 pandemic. Firstly, the study indicated the trend of the spread and symptoms of COVID-19, which was divided into three stages. Secondly, the results of the sentiment analysis and emotional tendency showed that the people had a negative outlook toward COVID-19. Thirdly, topic modeling and themes relating to COVID-19 and the outbreak were divided into three categories, including (1) emergency of COVID-19 impact, (2) the epidemic situation and how to control it, and (3) news and social media reporting on the epidemic.

Conclusions: Sentiment analysis and topic modeling can produce useful information about the trend of COVID-19 pandemic and alternative perspectives to investigate the COVID-19 crisis which has created considerable public awareness around the world. This finding shows that Twitter is a good communication channel for understanding both public concern and awareness about COVID-19 disease. These findings can help health departments to communicate information as to what the public thinks about the disease.

Keywords: COVID-19; Twitter; Social media; Infoveillance data; Health informatics

Introduction:

In the context of infectious disease, there have been many outbreaks in the human population causing damage to both lives and economy [1], such as swine flu in 2009, which originated in Mexico from influenza virus in swine. Swine flu caused several illnesses and the American government later announced that swine flu was a public health emergency. Other outbreaks of diseases have continuously occurred every 5-10 years, such as Ebola [2], SARS [3] and Middle East respiratory

syndrome (MERS) [4]. At the end of 2019, a significant respiratory disease outbreak happened again, this time originating in Wuhan, China. The World Health Organization (WHO) reported a cluster of cases of pneumonia in Wuhan on December 31, 2019. After that, the disease was defined by WHO as coronavirus disease 2019 (COVID-19). COVID-19 is a new infectious disease spread by respiratory droplets and contact and is generally infectious to human beings. COVID-19 has had an unprecedented impact on the world, with more than 9,000,000 confirmed cases and 500,000 reported deaths in more than 200 countries worldwide. On January 30, 2020, WHO reported that the COVID-19 was a Public Health Emergency of International Concern (PHEIC).

Social media platforms have been used to predict and understand the characteristics and status of disease outbreaks through the use of Google searches, Facebook or YouTube, with rich and useful information that can predict and explain the outbreak [5]. Text mining is also used to extract health information from social media comments, such as on Twitter [6]. Twitter data allows researchers to obtain large samples of user-generated content, thereby garnering insights to inform early response strategies. Social media data text mining has been used in order to track diseases. It is also used to assess public knowledge concerning health issues, leading to disease forecasting. Text analysis with Twitter data is one of the most important areas of focus in medical informatics research [7].

The coronavirus disease (COVID-19) is a scientifically and medically novel disease which is not fully understood as it needs to be consistently and deeply studied. In the past, research on the COVID-19 outbreak was only able to predict quantity data such as the number of outbreaks, not infoveillance data [8-9]. Most of the data reported is epidemiological, for example data from medical units or scientific laboratories. This can be challenging, as in the initial stage of the outbreak most data are incomplete due to inadequate diagnostic and testing capabilities, with incomplete epidemiological data regarding confirmed cases. The use of social media information to analyze syndromic surveillance focusing on public health-related concerns utilizing online information and content is essential. One important reason is that, during an outbreak, social media plays a critical role as a platform reflecting real-time public panic through comments. Twitter, one of these social media platforms, is often used as the main communication device when there is a disease outbreak. It can be used as a tool to keep an eye on an outbreak in a wide area, especially in severe situations. Twitter also provides rich information and outbreak patterns as well as the locations of outbreaks. This is very useful to provide insight regarding the disease and outbreak issues.

For COVID-19, there is still a lack of research studies on the spread of the disease based on deep information from social media to study the behavioral perspectives of online users as well as emergent conversations on COVID-19. Recently, in 2020, some research studies have been performed, such as the one by Shen et al. [10] on mentions of symptoms and diseases on social media to predict COVID-19 or by Huang et al. [11] on the characteristics of COVID-19 patients in China. However, these studies focused mainly on China, not other countries. Abd-Alrazaq et al. [12] conducted an infoveillance study on aspects of the COVID-19 pandemic, aiming to study the main topics related to the disease. However, studies are missing on themes and sentiment analysis on the timelines and trends of COVID-19 symptoms at the beginning of the outbreak.

Twitter data depict a more general population and serve as a leading indicator of COVID-19 since this platform includes tweets from healthy users who experience symptoms that might be treatable. Thus, policymakers may be missing valuable insights by excluding Twitter from COVID-19 information sources. In this study, user tweets showed responses to the crisis. This study used text-mining method to understand public perceptions of COVID-19 and to uncover meaningful themes of concern during the beginning of the outbreak in China and its spread throughout the world. The study focuses on investigating the behavioral perspectives of online users and emergent

conversations around COVID-19. The objective was to answer two research questions. Firstly, sentiment analysis was used to understand the themes underlying public perception in terms of sentiments and emotions towards COVID-19. Secondly, topic modeling was used to define the emergent themes and discourse regarding COVID-19.

Methods:

Data Collection:

The objective of this study was to answer questions relating to themes, public concerns and sentiments regarding the COVID-19 outbreak through social media analytics. The data were collected from Twitter in order to build a database on the COVID-19 epidemic pattern. The data from Twitter is acceptable for research, with high usefulness and richness [13]. Twitter is a medium for million people to express their views on any issue or topic. For example, during previous events such as natural disasters or disease outbreaks, people have used Twitter to express their feelings [14-15], particularly when there is a global epidemic such as the case of COVID-19. This study collected all tweets using the Twitter Streaming API, which is a JAVA application to connect to Twitter Streaming and store the raw data in a MySQL database. Twitter streaming API allows near real-time access to the global stream of public tweets that match the specified keywords. The tweet database has to formulate specifying keywords and metadata such as language, source, data range and location. This search used keywords and specific hashtags (#) such as “coronavirus”, “covid_19”, “2019-nCov” and “covid-19” in English language by “searchtweets” *(https://ptpi.org/project/searchtweets/)*. The scope of the keywords and specific hashtags determined what tweets will be delivered on the stream. This study employed a purposive sampling approach from active Twitter users published between December 13 and March 9, 2020. We selected the 1,000 most recent tweets per day representing conversation activities about COVID-19 on Twitter. This process was repeated to collect a total of 107,990 tweets during this time period. The main objective was to answer the research questions between the end of 2019 and the beginning of 2020 during which the outbreak started in China, then spread to Europe and America. This particular period is important regarding public concerns in relation to the COVID-19 outbreak.

Processing and Data Analysis:

To start data processing, the tweet texts were managed by a series of functions to remove URLS, emojis, special characters, retweets, # or hyperlinks pointing to websites and excluded the occurrence of related diseases contaminating the results as much as possible. Stop words in English were also removed, as well as words like “corona” or “virus” relating to other topics. There were three steps of data preparation, which were sampling, data collection and pre-processing of raw data. The tweets were then converted into a corpus (text mining structure), a document-term matrix and a used term frequency-inverse document frequency (TF-IDF), which is a numerical statistic used to reflect the importance of a word in a corpus. To obtain the output of the scenario, tweet data were analyzed; in order to extract the tweet data, “Twitter API” was used. API needed to be signed up on Twitter and also had to login into the developer Twitter account. The following step was an application or an API that needed to be developed, which can be used to provide the keys and the tokens for using it in the programming environment.

Data analysis was not only focused on COVID-19 in the overall picture but also scoped down the analysis based on the keyword and specific hashtag (#) such as “Symptoms”, “Outbreak” and “Pandemic”. There were three types of analysis in order to answer both research questions, as mentioned earlier. Firstly, the data analysis was focused on word frequencies in the corpus of the text

mining structure and visualized through word clouds to display the most common topics. The analysis also included time series using “retweet_count” and “favorite_count” as a proxy of social media activity for the activity intensity trend in Twitter relating to COVID-19 in order to see trends and timelines. Secondly, sentiment analysis, a natural language processing (NLP) approach, was used to categorize the sentiment appearing in Twitter messages. This involved the use of keywords appearing in the search topics and explored the sentiment analysis of each search topic related to COVID-19, including word frequency statistics and word clouds.

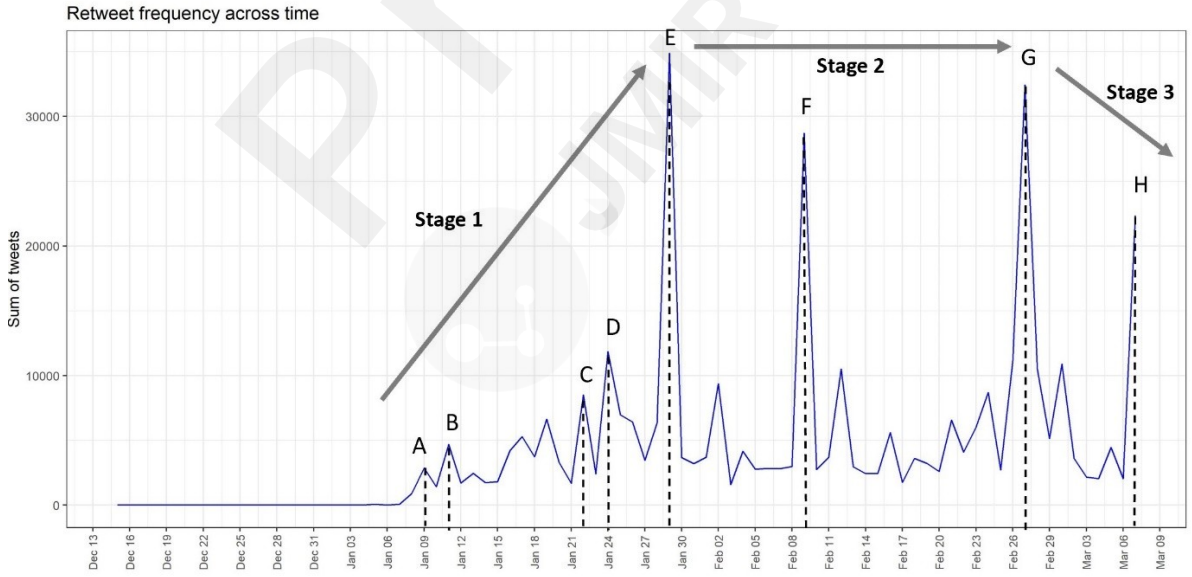
Finally, topic modeling was used, i.e. unsupervised machine learning analysis, to identify the most common topics in the tweets as well as to categorize clusters and find themes from keyword analysis. In order to perform topic modeling, the latent Dirichlet allocation (LDA) algorithm was applied. The main target was to map the given documents to the set of topics so that the words are extracted by those topics. LDA is a widely used topic modeling algorithm.

Results:

Search Trend on the COVID-19 Pandemic:

From the English Twitter messages, the analysis of COVID-19 epidemic trends did not aim to see the volume of daily tweet messages. The volume of tweet analysis was not relevant in this instance, as explained in the Methods section, as this study only downloaded 1,000 tweets per day (due to the restrictions placed on us by Twitter) with the stipulated keywords. The values such as average, min and max were not meaningful statistics in this instance. The intent here was to measure the intensity of Twitter activities relating to COVID-19; this study made use of the sum total of retweets as a proxy of the intensity of activity in Twitter.

Figure 1: Trend of the COVID-19 pandemic



Point A = Novel coronavirus isolated
Point B = First fatal case reported
Point C = The first confirmed American case of COVID-19 was declared
Point D = 835 cases reported in China
Point E (Peak 1) = WHO declared a public health emergency of international concern
Point F (Peak 2) = WHO announced name “COVID-19”

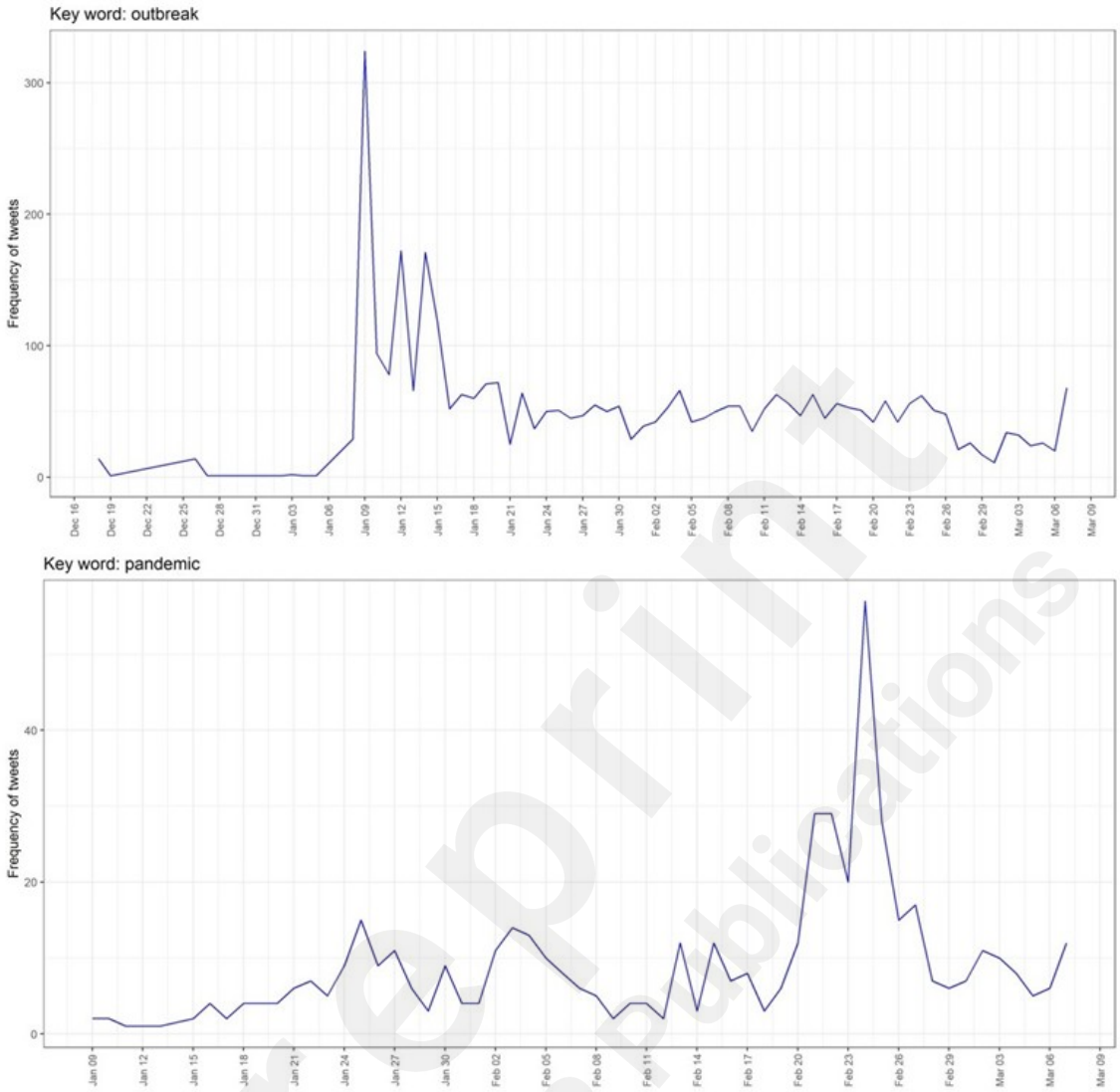
Point G (Peak 3) = Europe and Italy saw a spike in infections
Point H (Peak 4) = The number of the persons affected by COVID-19 surpassed 100,00

Figure 1 illustrates the retweet frequency, showing that the trendline increased from January 7-9 and the first peak with the highest intensity was on January 28-29 (Point E), with a second peak on February 9-11 (Point F) and a third peak on February 27-28 (Point G). The fourth peak was on March 6 (Point H). This result indicates public perceptions focusing on COVID-19 by Twitter intensity in the first period during January to its peak at the end of the month. It could be said that this showed an incubation period or early stage (Stage 1) when first-hand data about the severity of the emerging COVID-19 outbreak, including the evidence of human-to-human transmission, started to appear. Data compilation on the words related to symptoms resulting from COVID-19 infection at the prodromal phase, including fever, dry cough and malaise, are non-specific. On January 20-22, the first confirmed American case of COVID-19 was declared in Seattle and infection in healthcare workers was occurring. The spread was more severe and become more general until the end of January, when the USA declared a public health emergency. That period of time was when the Tweet message intensity was at its first peak.

Then, this developed into the worldwide epidemic period (Stage 2), with an epidemic representing wider spread with an effect on worldwide health and economic activity. It is now known that, during the epidemic period, the outbreak significantly moved out of China to other countries, including Hong Kong, Taiwan and Macau, as well as East Asian countries such as Japan and South Korea. During this time, there was talk about the fatality rate, which was as high as 3%, until the panic or peak period in the end of January. The second peak occurred on February 7 when WHO officials announced that they had identified a new virus called 2019-nCoV (COVID-19), causing high Twitter activity, leading to intensity. On February 27 as the third peak, COVID-19 reached Europe and Italy saw a spike in infections, which jumped to 650. It can be said that these two events indicated a complete pandemic as COVID-19 had spread all over the world from Asia to America and Europe, with the pandemic (worldwide epidemic) starting in Italy. Later, there was preparation following social distancing and lockdown. This was a stable stage (Stage 3) in the aspect of public perception. Even though the Twitter intensity was high again on March 6, when the number of the persons affected by COVID-19 surpassed 100,000. COVID-19 continues to spread, even as the WHO urges countries to “do more”.

Figure 2 shows the Twitter intensity for two keywords, i.e. “outbreak” and “pandemic”, which have different meanings. The peak of the trendline for the keyword “outbreak” was on January 9-11. As the infections increased, Chinese officials said that they had identified a new virus from the coronavirus family. It was named 2019-nCoV (COVID-19). This was the beginning of the outbreak of COVID-19, represented by a high activities of tweet messages including the word “outbreak”. Before reaching the pandemic level, the trendline showed a peak on February 24 as the virus spread worldwide from Asia to other continents. During that time, WHO announced worldwide epidemics, affecting countries in different ways. Using the word pandemic now fits the facts.

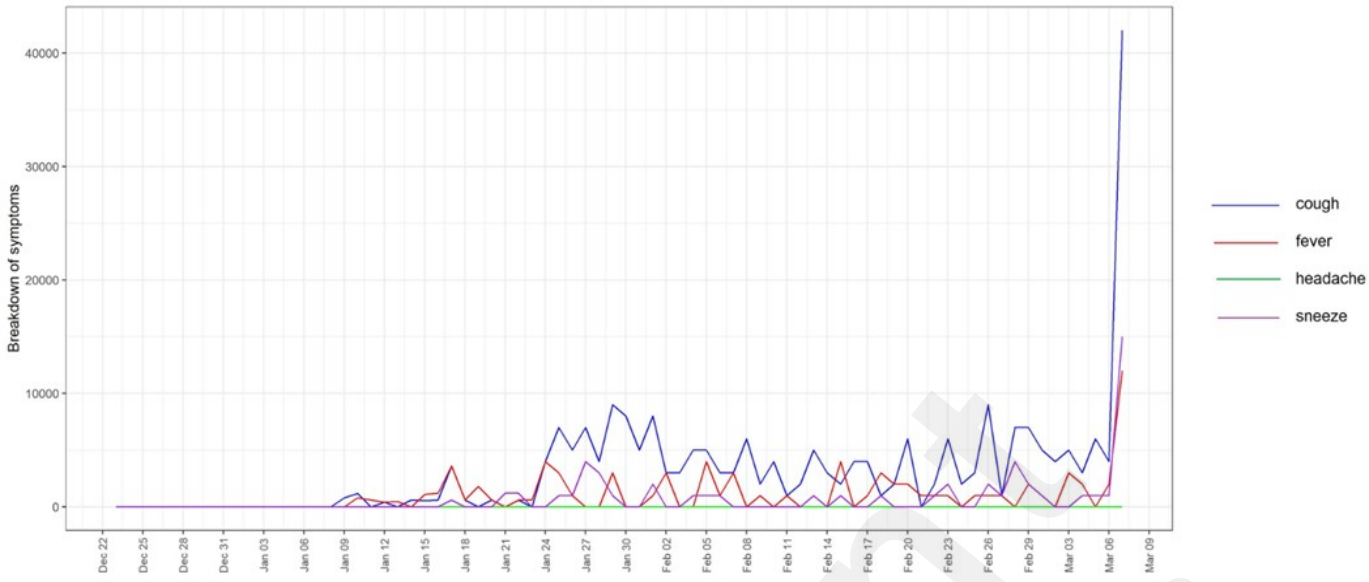
Figure 2: Twitter intensity for both keywords, including “outbreak” and “pandemic”



Search Trends on COVID-19 Symptoms:

Content analysis can be used to analyze words or messages showing that things happened after some incident or for the study of symptoms. Content analysis is used to inspect information or content created or written symbols. Word frequency count is another method widely used in content analysis. Figure 3 shows the word frequency count as trendlines regarding the symptoms of COVID-19 through Twitter information and may reflect views and concerns about COVID-19 symptoms. The two key symptoms of COVID-19 are cough and fever [16], as well as general symptoms such as headache and sneezing. However, the rest of the symptoms were negligible (e.g. body pain, runny nose, skin rash, frequent urination) as only the main symptoms were required to plot the graph and provide a clear analysis. The reason why the word “pneumonia” was removed because this condition describes inflammation of the tissue in one or both lungs. The study removed pneumonia in order to look at other related symptoms and rank the Twitter data mentions daily to indicate public awareness as well as the trend of COVID-19 symptoms.

Figure 3: Trend of COVID-19 symptoms

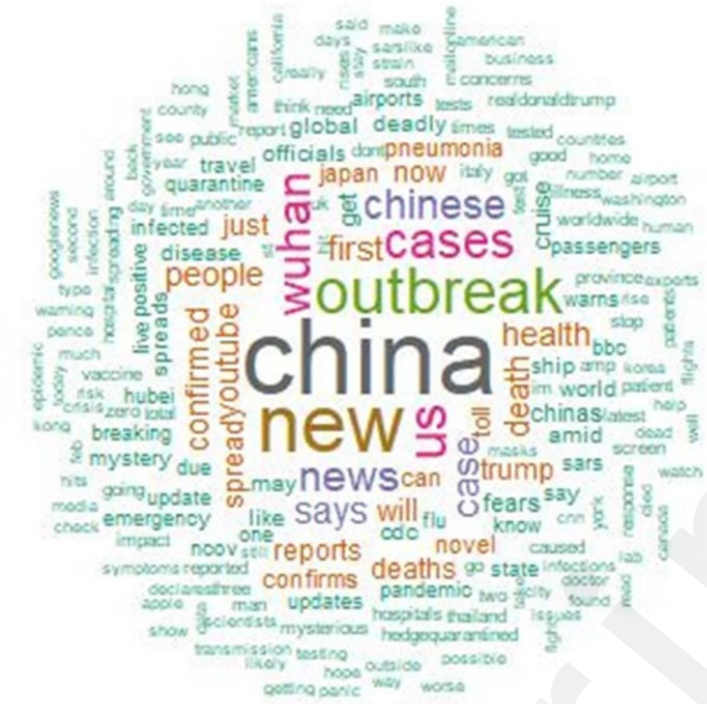


In Figure 3, the timelines of tweets mentioning symptoms of COVID-19 is presented. The analysis extracted messages mentioning at least one of the symptoms in the list. Before 24 January, coughing was not mentioned as much as fever, followed by fever and sneeze. Headache had the lowest word frequency count. After January 24, coughing became a clear symptom with the most mentions, then fever with sneezing, while headache was rarely mentioned with no change in the trend as time passed. It can be said that fever might be an early stage symptom, leading to coughing, and headache might be the next symptom with the fewest mentions. An interesting point was on March 6, which was the peak of coughing and fever as this was during the pandemic period with the highest frequency of mentions.

Frequency of keywords related to COVID-19:

This analysis used word clouds that can represent the visualization of tweet texts. It highlights words according to frequency. Word clouds of frequent words provided deeper level insights in tweets related to COVID-19 posted by people in this study. According to Figure 4, the words appearing were related to “China”, showing the origin of COVID-19. Moreover, the word “new” showed the spread of a new virus, including the word “outbreak” that also reflected the spread as a continuous epidemic. The secondary words in the word cloud were “Wuhan”, “death”, “health”, “people”, “spread”, and “confirmed”, as they depicted the nature and perspective of the population towards COVID-19 disease.

Figure 4: Word cloud of frequent keywords related to COVID-19



When specifically considering the frequency of keywords in the specific search, we used the words “outbreak” and “pandemic” to look at the different perspectives of different types of spread. Theoretically, outbreak means a greater-than-anticipated increase in the number of endemic cases. It could be a single case in a new area. If it is not well controlled, an outbreak can develop to be an epidemic. In this perspective, the words with the highest frequency for outbreak were “China” and “Wuhan” (Figure 5a), showing the first country/city to report the outbreak. Other words relating to COVID-19 were “pneumonia”, which describes infected lungs with COVID-19. It was the pilot symptom mentioned in the outbreak period when there were a number of pneumonia patients in Wuhan, China, leading the public health ministry to declare a new epidemic. The officials tried to find the cause and the infected, and attempted to control the outbreak area. Moreover, there were other words such as “disease”, “new”, “death” and “mystery”, showing the point of view during the outbreak, i.e. the period when COVID-19 had not reached all over the world; this analysis was restricted to only the beginning of COVID-19, which was characterized by the mention of pneumonia in China.

Figure 5: Word cloud of frequent keywords related to COVID-19 epidemic (5a) and pandemic (5b)

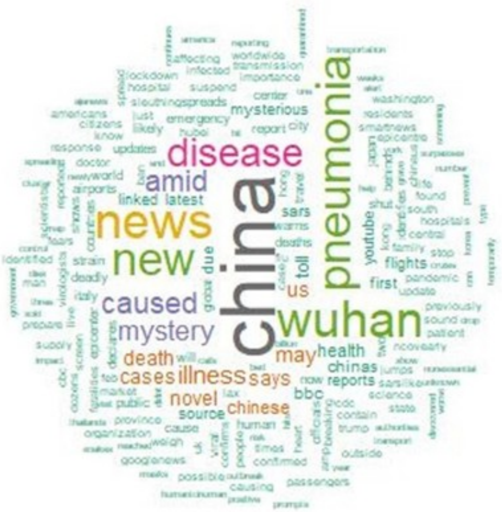


Figure 5a

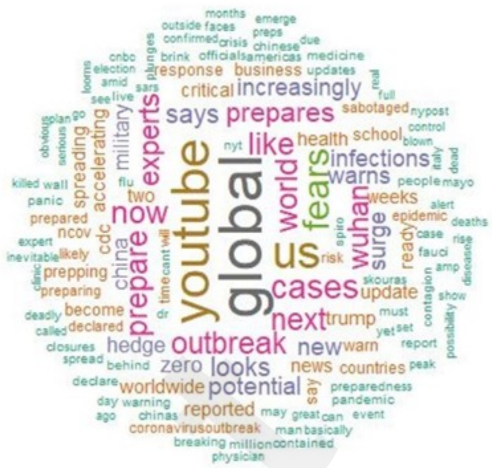


Figure 5b

Frequency keyword searching was performed using the word “pandemic”, which according to the definition refers to a disease that affects a large number of people within a community, population or region that is spread over multiple countries or continents. This meaning reflects the frequency of perceptions towards COVID-19 in the context of a pandemic. The words with the highest frequency were “global”, “world”, “outbreak” and “YouTube” (Figure 5b), which were mentioned with the word pandemic. There was an increase in the number of cases in locations around the world. As a social media platform with billions of daily views, YouTube has tremendous potential to both support and hinder public health efforts regarding COVID-19 information. Social media was mentioned as the media of pandemic news trends. Moreover, other related words were “expert”, “prepare” and “now”, showing an awakening towards the pandemic and the preparation and experts needed to stop this pandemic. The WHO declared COVID-19 a pandemic due to its spread all over the world as well as the domestic connection, which normally happens after the arrival of passengers from pandemic countries, leading to the infection of family members or friends and spread in new, hard-to-control areas.

Sentiment analysis on COVID-19:

Sentiment-level analysis further enriched the analysis through the clear identification of negative and positive topics on COVID-19. From the analysis of sentiments, it was found that 22.12% of the tweets contained positive sentiments as compared to 77.88% negative sentiments (Figure 6). This signifies that people had a negative outlook towards COVID-19. According to Figure 7, on January 6-9 when there was an official announcement on the discovery of coronavirus (COVID-19), negative sentiment went higher than positive sentiment. During the period with a high death rate and pandemic level outbreak up to March 6, the negative sentiment was high again when the number of the persons affected by COVID-19 crossed 100,000. COVID-19 continued to spread even as the WHO urged countries to “do more”, as has been mentioned earlier.

Figure 6: Sentiment wheel of tweets

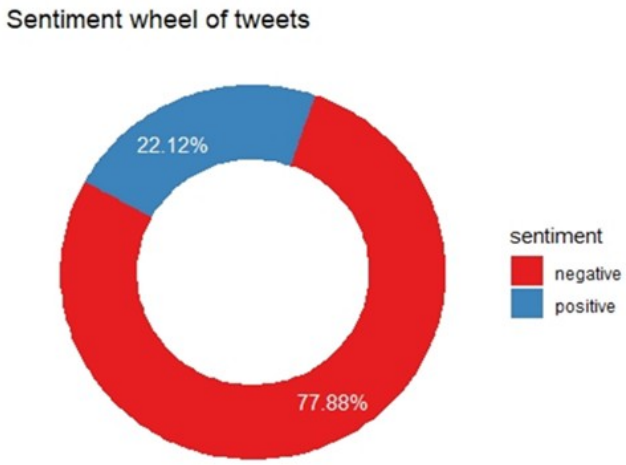


Figure 7: Sentiment analysis on COVID-19

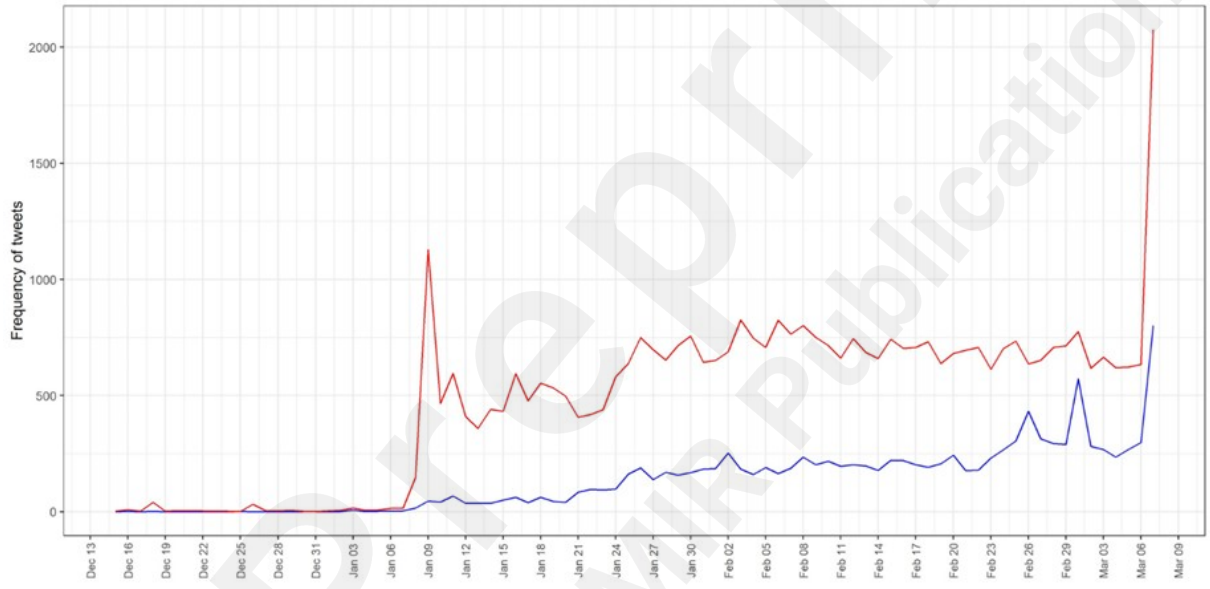
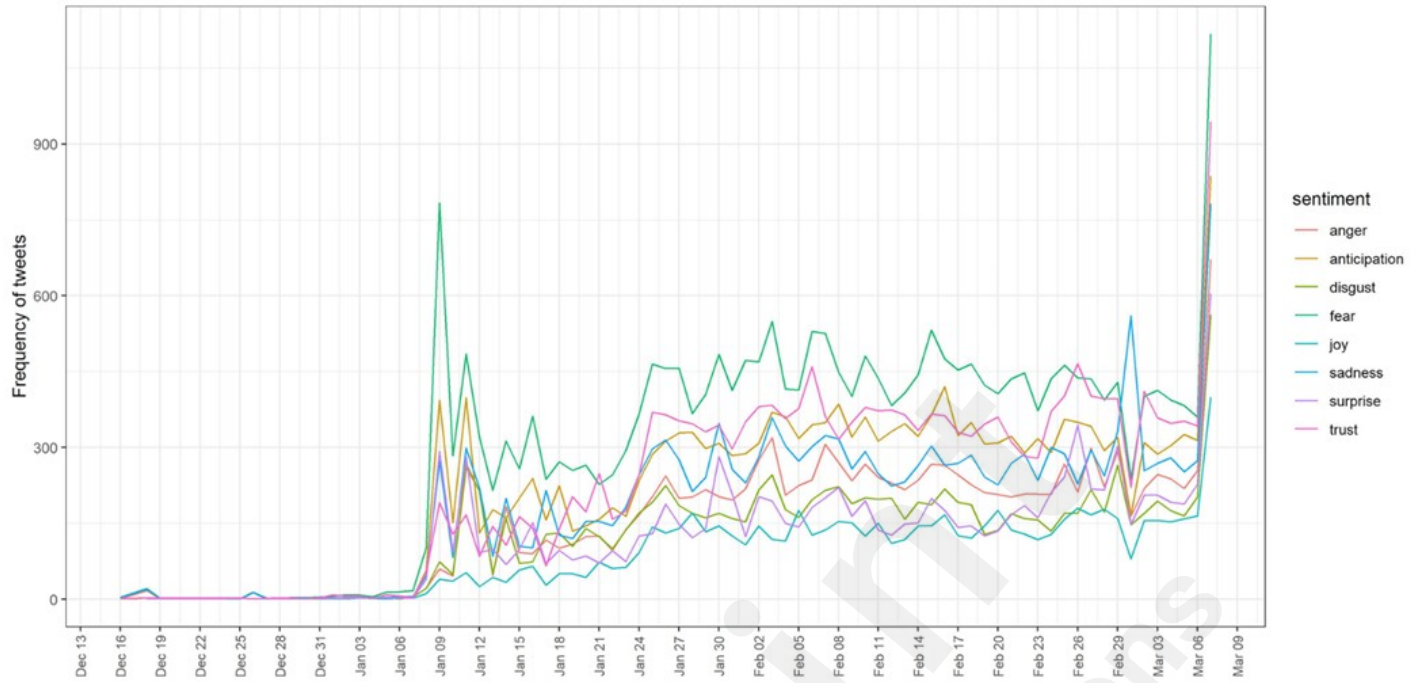


Figure 8: Sentiment analysis using the NRC sentiment lexicon



For the more detailed analysis of the tweets, the emotion quotients associated with the tweets were analyzed (Figure 8). Sentiment analysis using the NRC sentiment lexicon allowed the researcher to express each document comprising 10 basic emotions: “anger”, “anticipation”, “disgust”, “fear”, “joy”, “negative”, “positive”, “sadness”, “surprise” and “trust” [23]. Additionally, the terms “positive” and “Negative” were removed as they did not bear positive or negative emotions, as would otherwise be indicated by the NRC sentiment lexicon. As a result, there were total of eight emotions evaluated in this analysis. Among the eight emotions, “trust” and “joy” were positive emotions, while “anger”, “sadness”, “fear”, and “disgust” were considered negative emotions. “Surprise” and “anticipation” can be either positive or negative depending on the context.

During the analysis of the emotional quotient of the tweets, it was found that over half of the tweets across the world were defined by three emotions, including “fear”, “trust” and “anticipation”. Figure 9 shows that the tweets with the “fear” emotion comprised about one fifth of the total tweets analyzed (21.19%). Following the “fear” emotion was the “trust” emotion, which indicated that people were looking forward to the recovery or solutions from experts. Similarly, the “anticipation” emotion was associated with almost 15.16% of the tweets, which again strengthens the positive sentiments of the people. Negative emotions such as “sadness”, “anger” and “disgust” were seen in a portion of the tweets, with shares of 13.20%, 10.73%, 8.86%, respectively. Only a small portion were categorized as “joy”, which is a positive emotion, with 6.18%. These results show that the people had a negative outlook toward COVID-19. As shown in Figure 10, the positive sentiment keywords were “patient”, “protect”, “tough”, “safe” and “cure”, while the main words in Twitter with negative sentiments were “outbreak”, “virus”, “death”, “infected” and “fear”.

Figure 9: The analysis of the emotional quotient of tweets

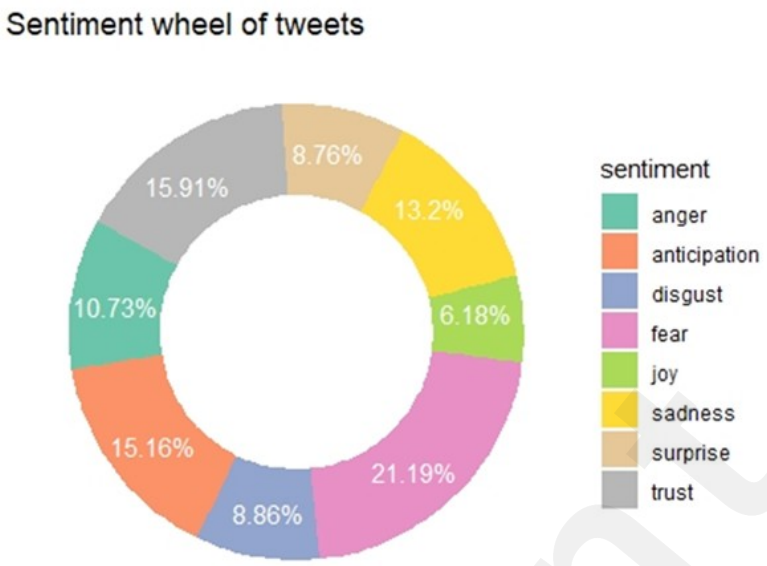
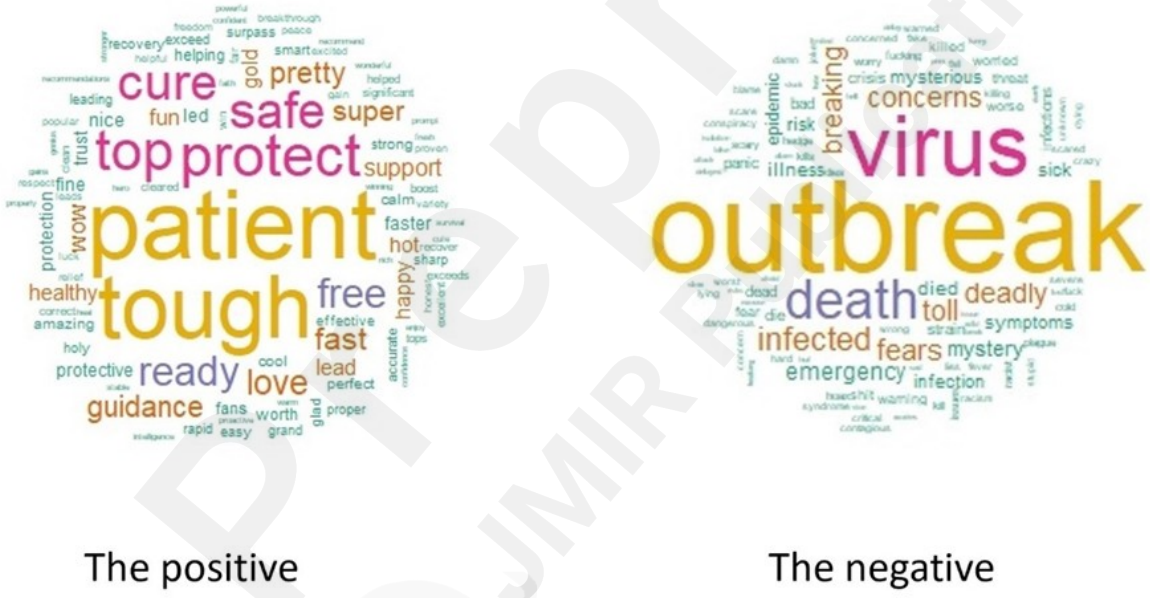


Figure 10: Word cloud of the positive and negative frequent keywords related to the COVID-19 epidemic



For the next phase of the sentiment analysis, a word cloud was created using the most emotions or used words, categorized by color. The sentiments were part of the sentiment library in R (‘textdata’), in which each emotion was separated, with no overlap. The word cloud was developed with the whole corpus of tweets. The results in Figure 11 show that the word “death” was tweeted most in this context. As illustrated in Figure 11, words such as “death”, “mystery (mysterious)”, “epidemic” and “guess” were used the most and related with the emotion of “surprise”. Words such as “pneumonia”, “flu”, “infection”, “panic” and “quarantine” were tweeted frequently and related to “fear”. Words such as “disease”, “deadly” and “sick” were tweeted frequently with “disgust”. Words like “pandemic”, “illness” and “hospital” were frequently tweeted with “sadness”. For the positive sentiments, people also tweeted words like “hope”, “safe” and “diamond” with the emotion of “joy” while words such as “confirmed”, “doctor”, and “expert” were regularly used with emotions of “trust”.

This result indicates that when the majority of the people thought about COVID-19 pandemics, they had a negative feeling. Most of them were surprised with the mysterious disease with no prior information about how to treat it and the possibility of death. In addition, when talking about symptoms such as pneumonia, flu or infection, they usually felt great fear.

Figure 11: Word cloud using the most emotions or used words (categorized by color)

Topic Modeling:

COVID-19 Related Themes:

In this section, the emergent topics and themes were developed and summarized in Table 1 using topic modeling, which is the unsupervised classification method of documents, similar to clustering on numeric data finding natural groups of items even when it is not certain what is being looked for. The latent Dirichlet allocation (LDA) algorithm is a particularly popular method for fitting a topic model. It treats each document as a mixture of topics, and each topic as a mixture of words. This process allows tweeted messages to overlap each other in terms of content, rather than separate into different groups. This study introduced the “tidy () method” originally from the broom package [17]. The “tidytext” package provided this method in order to extract the per-topic-per-word probabilities, called beta, from the model.

Table 1: The emergent topics and themes regarding COVID-19

Category	Top 10 words	Topics
COVID-19 Keyword	China, death, first, new, outbreak, pneumonia, reports, spread, toll, Wuhan	**Topic 1**: China reports new cases and deaths of coronavirus outbreak in China with deadly pneumonia.
	Cases, going, just, like, nCov, positive, says, tested, tests, US	**Topic 2**: The epidemic situation and confirmed cases of coronavirus disease.
	Outbreak, people, will, news, Wuhan, Chinese, get, knows, disease, can	**Topic 3**: People know about coronavirus disease and outbreak on news
	US, Trump, Chinese, cruise, japan, passenger, spread, spreads, ship, CDC	**Topic 4**: The spread of the coronavirus and how to control the disease from overseas to US for fellow passengers with confirmed cases.
	Case, health, now, first, confirmed, China, fear, global, emergency	**Topic 5**: Declare health concerns and fear on coronavirus cases as emergency worldwide.
	Amp, Wuhan, good, UK, YouTube, second, days, apple, news, article,	**Topic 6**: News and information reports in social media on the epidemic.

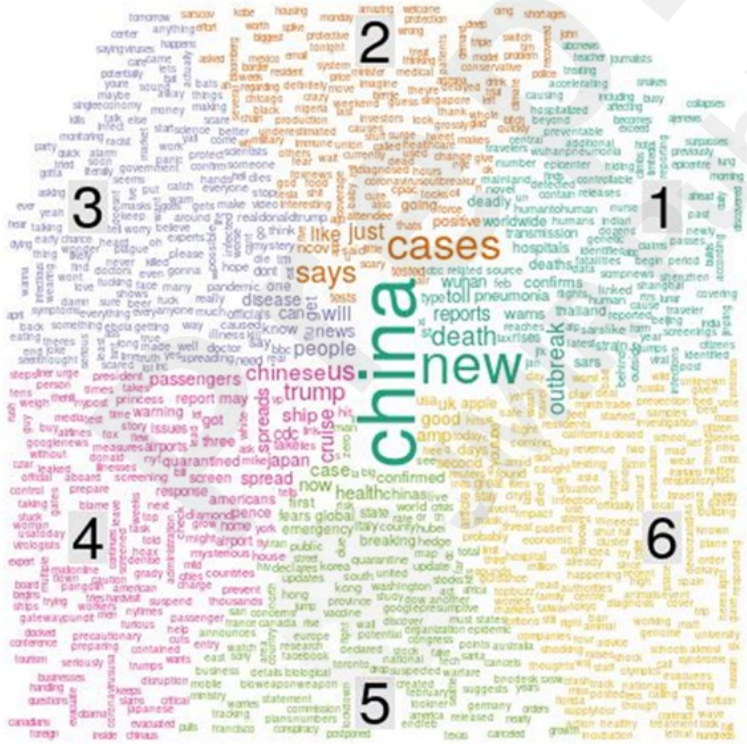
Theme 1 = New cases and the COVID-19 pandemic (Topics 1 and 5)
Theme 2 = The epidemic and how to control it (Topics 2 and 4)
Theme 3 = News and social media reports on the epidemic (Topics 3 and 6)

The objective of topic modeling was to answer the research question: What are the emergent topical themes and discourse regarding COVID-19 using topic modeling? In the first set, the emergent topics were identified from 10 of the highest key words in each group as topics. Then, each topic was concluded together as a theme. The topics and keywords were identified from a combination of clusters and word clouds which represented a collection of posts that were classified in the Tweeter. Figure 12 depicts the word cloud from the topic categories where the size of each word is proportional to the density p(word|topic).

The per-topic-per-word probabilities produced by LDA by extracting the beta matrix. As shown in Figure 13, the ten most common words in each topic by beta were considered in the study. This study used these words to provide each topic with a degree of semantic interpretation in the related contexts through relevant topic descriptions. The higher the beta value is, the greater the possibility of a relatable word appearing in the category. By this approach, six topics were classified based on the per-topic-per-word probabilities (beta) as follows. Topic 1 involved discussion related to China reporting new cases and the outbreak of COVID-19 with deadly pneumonia. Examples of keywords included “China”, “death”, “first”, “new”, “outbreak”, “pneumonia”, “reports”, “spread”, “toll” and “Wuhan”. The highest beta value was the world “China”. Topic 2 involved the epidemic situation and confirmed positive cases of COVID-19. The examples of keywords included “cases”, “going”, “just”, “like”, “nCov”, “positive”, “says”, “tested”, “tests” and “US”. The words in Topic 3 collectively related to what people knew about COVID-19 and the outbreak on the news, which were terms such as “outbreak”, “news”, “know” and “disease”. Top words in Topic 4 described the spread of COVID-19 and how to control the disease, while those in Topic 5 captured health concerns and fear regarding COVID-19 as an emergency. Some of the top 10 words were “wear”, “emergency”

and “health”, generally believed to be terms related to health concerns. Topic 6 was collectively associated with news and information reports on social media regarding COVID-19. The top terms in Topic 6 were “news”, “articles” and “YouTube”.

Figure 12: Word cloud from topic categories

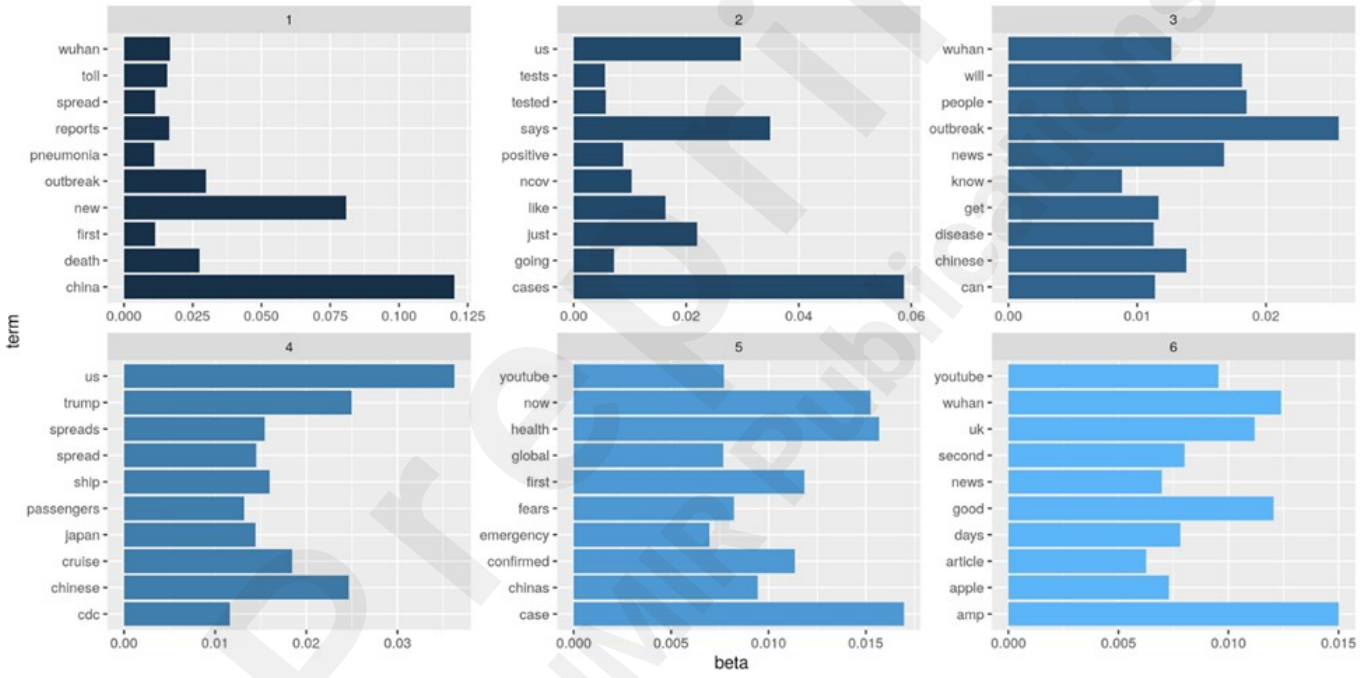


The qualitative content analysis approach allowed the users to categorize these topics into different distinct themes. As shown in Table 1, tweet samples in each topic were categorized and identified six topics in different themes. A theme identified based on keywords in Topic 1 and Topic 5 was about “New cases and the pandemic of COVID-19”, which provided information on the reporting of new cases and deaths in the COVID-19 outbreak and health concerns regarding a worldwide emergency. For example, the sample tweets reflected on “Top WHO official warned the world may be

‘dangerously unprepared’ for next pandemic as coronavirus outbreak spreads” or “The US health department declared the coronavirus a health emergency, 8 cases confirmed in the US, 259 dead over 11K infected in China”.

Also, the epidemic situation and how to control it theme was composed of Topic 2 and Topic 4, relating to the epidemic situation and confirmed cases of COVID-19 and its spread. Sample tweets such as “Fox News’ Maria Bartiromo predicted ‘hundreds of thousands of US coronavirus cases: ‘I don’t want to panic anybody’” or “As the coronavirus grows and infects and kills more people, Trump slashed the budget for the CDC that controls disease”. The final theme on news and social media reports on the epidemic was composed of Topic 3 and Topic 6, and was about the channels receiving news and information of COVID-19 such as “#Coronavirus has been dominating the news, but how much do we need to worry about it” or “China spent the crucial first days of the Wuhan coronavirus outbreak arresting people who posted”.

Figure 13: The per-topic-per-word probabilities produced by LDA by extracting the beta matrix



COVID-19 Outbreak Related Themes:

To explore the themes reflected by topics related to the COVID-19 outbreak on the keywords, this study used word clouds and topic modeling techniques to generate themes and co-occurrence topic keywords related to the COVID-19 outbreak. The results are shown in Figure 14. The main topic of public concern about the COVID-19 outbreak was the COVID-19 illness, and the public was extremely concerned about the status of the outbreak in Wuhan, China and the situation in the news. The high-frequency keywords of topics can be divided into three clusters.

Figure 14: Word cloud and topic modeling related to the COVID-19 outbreak

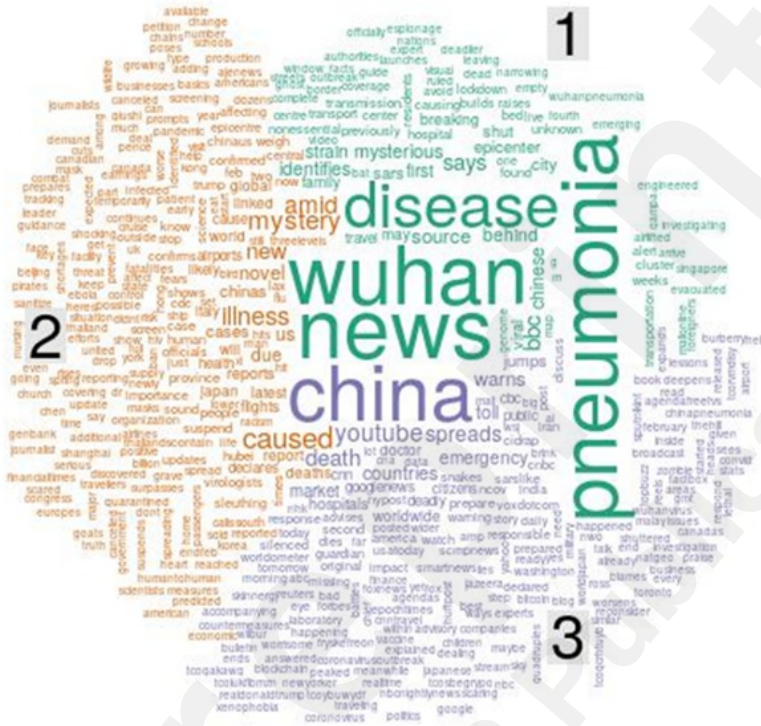


Table 2: The emergent topics and themes with “Outbreak” keyword

Category	Related words	Topics	Themes
Outbreak Keyword	Wuhan, pneumonia, identifies, strain	**Topic 1**: New strain of pneumonia identified in Wuhan	New mysterious illness, new strain of pneumonia caused by virus identified in Wuhan
	Novel, mystery, caused, illness	**Topic 2**: New mysterious illness caused by virus	
	China, spreads, jumps, warns, toll	**Topic 3**: Death toll could jump, China warns	China warns that death toll may jump

As shown in Table 2, Topic 1 captured discussions regarding the new strain of pneumonia identified in Wuhan. Examples of keywords included “news,” “Wuhan,” “pneumonia” and “disease”. Topic 2 involved searches for the new mysterious illness caused by virus. Examples of keywords included “new”, “cause” and “mystery”. Topic 3 included searches for the death toll in China. Examples of keywords in this topic included “China”, “death” and “emergency”. The theme based on Topic 1 and Topic 2 is about a new mysterious illness and new strain of pneumonia caused by a virus identified in Wuhan. This study also identified another theme, which was the China warned that the death toll could jump, based on Topic 3.

Discussion:

Principal Results:

During public health outbreaks, analyzing social media text can provide information with the useful insights about public concern, awareness or other perception [18]. Especially in the case of a new pandemic such as COVID-19, most research results and information are in the form of numbers or medical/scientific fact. The use of social media, especially Twitter, was useful for an explanation of the situation. This study found that social media can be used to measure public attention toward the COVID-19 epidemic [19].

This research attempted to understand the COVID-19 pandemic by gathering data from Twitter from December 2019 to March 2020, which was the beginning of the pandemic as it spread across the world. The results indicated three main aspects of public awareness and concerns which were (1) the trend of the spread and symptoms of COVID-19 in different stages, (2) sentiment analysis and (3) topic modeling and themes related to the COVID-19 outbreak.

The Twitter data analysis can be used to explain trends of the COVID-19 epidemic, shown in the trendline of the COVID-19 epidemic. It was divided into three main phases. The early or incubation stage was the phase in which the severity and the spread of coronavirus started to increase. The public started to be aware of its severity and rapid spread, then started to fear, especially in January when WHO announced a new virus related to pneumonia. The beginning of stage 1 was statistically significant and in accordance with sentiments of the people. It appeared that the negative emotions towards COVID-19 were very high in stage 1, as there was little information on the new virus. It increased until it reached the highest point at the end of January. This result is in accordance with a previous study [19] explaining the different stages of the public’s attention to the COVID-19 epidemic in China. There was a need to not undermine the possibility of a serious outbreak during the pre-crisis period [20].

The second stage was the worldwide epidemic stage. This period was characterized by the official announcement of the COVID-19 pandemic by the WHO as it spread across the world. In the third stage, as the number of the confirmed cases continued to increase, the public started to be aware and there was more scientific and medical understanding. There was preparation following social distancing and lockdown. This was a stable stage in the aspect of public perception. In the analysis of Twitter messages, the recognition of COVID-19 symptoms concluded that fever was the major symptom of COVID-19, in accordance with research results stating that fever was seen in 94.3% of cases and was the most common symptom present at the onset of illness (87.1%), followed by cough (36.5%) and fatigue (15.7%) [21-22]. The common symptoms remained consistent across several studies, including fever and cough [23]. Fever is understood to be a precursory indicator of COVID-

19. After that, the virus progresses to respiratory system, causing pneumonia and a severe cough [24]. Coughing was a significant symptom in the late stage of fever. However, there were other symptoms such as a stuffy nose and headache.

The result of the sentiment analysis of COVID-19 showed that the most important keyword was “outbreak”, related to the starting point of the disease in Wuhan, China. When it entered the pandemic phase in March, message perception in Twitter showed words such as “global”, “prepare”, “cases” and “YouTube”. The results depicted the development of the pandemic in limited areas, such as a city or country, to the world. The keyword “publicly aware” was different according to the spread stage including public emotion that was mostly negative over positive; “fear” was the most negative word [25-26]. In previous disease pandemics, negative sentiments were generally prevalent in social media [27]. A study by Raamkumar [28] also showed the research results from Facebook in several countries. The word “fear” related to COVID-19 was the most negative sentiment.

According to the news, the spread of COVID-19 was still mysterious with a chance of death, without no vaccine or treatment, resulting in public concerns. As the epidemic progressed, however, the public sentiment tended to be positive because more news was being reported at this stage. The public’s positive emotions increased. More information was available about prevention or protection, which is favorable to public health communication and promotion. The analysis showed positive feelings through keywords such as “trust”, “protect” and “safe” and demonstrated that the public still trusted experts and departments that would help them get through this situation, along with medical and scientific personnel who provided rapid and clear information to the public. The result was in accordance with previous research showing that people’s interest was related to the latest news and major events in infectious diseases on social media. Studies have also indicated that people would pay attention to and search for disease-related words as the spread of infectious disease changes [29].

The COVID-19 crisis has stimulated great public concern around the world. Topic modeling offers an alternative perspective to investigate the COVID-19 crisis. The research results from users on Twitter between December 2019 and March 2020 divided public concerns into six topics which were (1) new cases and deaths from the outbreak in China, (2) the epidemic situation, (3) people learning about the outbreak on the news, (4) the spread and how to control the disease from overseas, (5) health concerns and fear as the progressed emergency worldwide, and (6) news and information reports on social media. The topic modeling analysis results showed that messages in Twitter demonstrated public concern in three main themes, which were (1) emergency COVID-19 impacts, (2) the epidemic situation and how to control it, and (3) news and social media reporting on the epidemic. This finding showed that Twitter was a good communication channel for both individuals and organizations to publicize COVID-19 symptoms and preventive measures [30]. Previous studies have also shown that prevention and control procedures, as well as medical treatment, were major themes during previous disease outbreaks [31-32]. In the meantime, Twitter was also a good channel to distribute information to the public regarding the number of confirmed cases, patients, spread status, prevention and control. Moreover, there was sharing of information to facilitate prevention and Twitter notification as the public connection [33].

Limitations:
This study contained several limitations. First, it is worth mentioning that this study used some of the keywords related to COVID-19 to investigate trends and frequencies of keywords. The selected keywords may have been incomplete. The keywords used in this study can be extended to cover the search of Twitter messages by combining keywords related to COVID-19 and its symptoms. Further research could aim to identify the most relevant set of keywords. Second, this research was performed in the early phase of the pandemic that then progressed across the world. There was a

limitation in the scope regarding public awareness of the total picture as well as the pandemic cycle, and a study on public concern after the mentioned period might provide useful results for comparison. Third, although LDA was advantageous in extracting hidden themes, the scientific quality of the themes should be further validated. Furthermore, researchers can play a greater role in extracting themes. However, they need to reduce bias, which might happen when finding topic themes using topic modeling. Lastly, it may be difficult to find the perfect sources of social media since the amount of information regarding COVID-19 was overwhelming. This research collected data only from Twitter; further research should use other resources such as mass media or other data sources from social media information.

In conclusion, this study indicated the themes underlying public perception in terms of sentiments and emotions towards COVID-19. Furthermore, the results also show the emergent topical themes and discourse regarding COVID-19 using topic modeling. This research may be useful for policy makers and the government sector to determine the policy for COVID-19 control. The recognition of public concern and awareness was advantageous to see how the public thought about the disease at a particular time. When the results are connected, a valuable healthcare resource can be established for a future plan.

Conflicts of Interest:

None declared.

Abbreviations

COVID-19: Coronavirus disease
LDA: latent Dirichlet allocation
MERS: Middle East respiratory syndrome
NLP: a natural language processing
PHEIC: Public Health Emergency of International Concern
TF-IDF: Term frequency-inverse document frequency
WHO: World Health Organization

References

1. Kilbourne, E. D. Influenza pandemics of the 20th century. Emerg Infect Dis 2006; 12(1): 9.
2. Jacob ST, Crozier I, Fischer WA, Hewlett A, Kraft CS, Vega MDL, et al. Ebola virus disease. Nat Rev Dis Primers 2020 Mar 20;6(1):13.
3. de Wit E, van Doremalen N, Falzarano D, Munster VJ. SARS and MERS: recent insights into emerging coronaviruses. Nat Rev Microbiol 2016 Aug;14(8):523-534.
4. Ahase E. Coronavirus covid-19 has killed more people than SARS and MERS combined, despite lower case fatality rate. Br Med J 2020 Mar 18;368:m641.
5. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009 Feb;457(7232):1012-4.
6. Jahanbin K, Rahmanian V. Using Twitter and web news mining to predict COVID-19 outbreak. Asian Pac J of Trop Med 2020; 13;1-3.
7. Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PLoS One 2010 Nov;5(11): e14118.
8. Zhao Y, Cheng S, Yu X, Xu H. Chinese public's attention to the COVID-19 epidemic on social media: observational descriptive study. J Med Internet Res 2020;22(5): e18825.
9. Budhwani H, Sun R. Creating COVID-19 stigma by referencing the novel coronavirus as the "Chinese virus" on Twitter: quantitative analysis of social media data. J Med Internet Res 2020;22(5):e19301.
10. Shen C, Chen A, Luo C, Zhang J, Feng B, Liao W. Using reports of symptoms and diagnoses on social media to predict COVID-19 case counts in mainland China: observational infoveillance study. J Med Internet Res 2020;22(5):e19421.
11. Huang C, Xu X, Cai Y, Ge Q, Zeng G, Li X, et al. Mining the characteristics of COVID-19 patients in China: analysis of social media posts. J Med Internet Res 2020;22(5):e19087.
12. Abd-Alrazaq A, Alhuwail D, Househ M, Hamdi M, Shah Z. Top concerns of tweeters during the COVID-19 pandemic: infoveillance study. J of Med Internet Res 2020;22(4):e19016.
13. Dai H, Deem MJ, Hao J. Geographic variations in electronic cigarette advertisements on Twitter in the United States. Int J Public Health 2017 May;62(4):479-487. [doi: 10.1007/s00038-016-0906-9] [Medline: 27742923]
14. Nair, MR, Ramya GR, Sivakumar PB. Usage and analysis of Twitter during 2015. Chennai flood towards disaster management. Procedia Comp Sci 2017;115:350-358.
15. Fu KW, Liang H, Saroha N, Tse ZTH, Ip P, Fung ICH. How people react to Zika virus outbreaks on Twitter? A computational content analysis. Am J Infect Control 2016;44(12):1700-1702.
16. Geldsetzer P. Use of rapid online surveys to assess people's perceptions during infectious disease outbreaks: a cross-sectional survey on COVID-19. J Med Internet Res 2020;22(4):e18790.
17. Robinson, D. 2017. Broom: Convert Statistical Analysis Objects into Tidy Data Frames
18. Gui X, Wang Y, Kou Y, Reynolds TL, Chen Y, Mei Q, et al. Understanding the patterns of health information dissemination on social media during the Zika outbreak. AMIA

Annu Symp Proc 2017;820-829.
19. Zhao Y, Cheng S, Yu X, Xu H. Chinese public's attention to the COVID-19 epidemic on social media: observational descriptive study. J Med Internet Res 2020;22(5):e18825.
20. Lwin M, Lu J, Sheldenkar A, Schulz P. Strategic uses of Facebook in Zika outbreak communication: implications for the crisis and emergency risk communication model. Int J Environ Res Public Health 2018 Sep 10;15(9):1974.
21. Chen J, Qi T, Liu L, Ling Y, Qian Z, Li T, Li F, Xu Q, Zhang Y, Xu S, Song Z. Clinical progression of patients with COVID-19 in Shanghai, China. Journal of Infection 2020 Mar 19.
22. Huang C, Xu X, Cai Y, Ge Q, Zeng G, Li X et al. Mining the characteristics of COVID-19 patients in China: analysis of social media posts. J Med Internet Res 2020;22(5):e19087.
23. Sarker A, Lakamana S, Hogg-Bremer W, Xie A, Al-Garadi MA, Yang YC. Self-reported COVID-19 symptoms on Twitter: An analysis and a research resource. medRxiv 2020 Jan 1.
24. Murray C, Mitchell L, Tuke J, Mackay M. Symptom extraction from the narratives of personal experiences with COVID-19 on Reddit. arXiv preprint arXiv:2005.10454 2020 May 21.
25. Dubey AD, Tripathi S. Analysing the sentiments towards work-from-home experience during COVID-19 pandemic. J Innov Manag 2020; 8:1.
26. Kleinberg B, van der Vegt I, Mozes M. Measuring emotions in the COVID-19 real world worry dataset. arXiv preprint arXiv:2004.04225. 2020 Apr 8.
27. Mamidi R, Miller M, Banerjee T, Romine W, Sheth A. Identifying key topics bearing negative sentiment on Twitter: insights concerning the 2015-2016 Zika epidemic. JMIR Public Health and Surveillance 2019;5(2): e11036.
28. Raamukar AS, Tan SG, Wee HL. Measuring the outreach efforts of public health authorities and the public response on Facebook during the COVID-19 pandemic in early 2020: cross-country comparison. J Med Internet Res 2020;22(5):e19334.
29. Li L, Zhang Q, Wang X, Zhang J, Wang T, Gao TL, Duan W, Tsoi KK, Wang FY. Characterizing the propagation of situational information in social media during covid-19 epidemic: A case study on weibo. IEEE Trans on Com Social Syst 2020 Mar 20;7(2):556-62.
30. Park HW, Park S, Chong M. Conversations and medical news frames on Twitter: Infodemiological study on COVID-19 in South Korea. J Med Internet Res 2020;22(5):e18897.
31. Liu Q, Zheng Z, Zheng J, Chen Q, Liu G, Chen S, Chu B, Zhu H, Akinwunmi B, Huang J, Zhang CJ. Health Communication Through News Media During the Early Stage of the COVID-19 Outbreak in China: Digital Topic Modeling Approach. J Med Internet Res 2020;22(4):e19118.
32. Shah K, Kamrai D, Mekala H, Mann B, Desai K, Patel RS. Focus on mental health during the coronavirus (COVID-19) pandemic: applying learnings from the past outbreaks. Cureus 2020 Mar;12(3).
33. Breland JY, Quintiliani LM, Schneider KL, May CN, Pagoto S. Social media as a tool to increase the impact of public health research. Am J Public Health 2017 Dec;107(12):1890-1891.

Supplementary Files