Twitter discussions and concerns about COVID-19 pandemic: Twitter data analysis using a machine learning approach (Preprint)

Preprint · May 2020

DOI: 10.2196/preprints.20550

Jia Xue
University of Toronto

Chengda Zheng
University of Toronto

Twitter discussions and concerns about COVID-19 pandemic: Twitter data analysis using a machine learning approach

Jia Xue, PhD
Factor-Inwentash Faculty of Social Work
& Faculty of Information
University of Toronto
jia.xue@utoronto.ca

Junxiang Chen, PhD
School of Medicine
University of Pittsburgh
Juc91@pitt.edu

Ran Hu
Factor-Inwentash Faculty of Social Work
University of Toronto
ranh.hu@mail.utoronto.ca

Chen Chen, PhD
Middleware system research group
University of Toronto
Chenchen@eecg.toronto.edu

Chengda Zheng
Faculty of Information
University of Toronto
chengda.zheng@mail.utoronto.ca

Tingshao Zhu*
Institute of Psychology
Chinese Academy of Sciences
tszhu@psych.ac.cn (T.S.Z.)

***Corresponding author**: Tingshao Zhu, Professor, 16 Lincui Road, Chaoyang District, Beijing 100101, China.

Acknowledgements: Publication of this article is supported by Funding by National Natural Science Foundation of China, grant number 31700984, and data collection is supported by Artificial Intelligence Lab for Justice at University of Toronto, Canada.

Twitter discussions and concerns about COVID-19 pandemic: Twitter data analysis using a machine learning approach

Abstract

Background: Public response to the COVID-19 pandemic is under measured (Stokes et al., 2020). Twitter data are an important source for the infodemiology study of public response monitoring.

Objective: The objective of the study is to examine coronavirus disease (COVID-19) related discussions, concerns, and sentiments that emerged from tweets posted by Twitter users.

Methods: We collected 22 million Twitter messages related to the COVID-19 pandemic using a list of 25 hashtags such as "coronavirus," "COVID-19," "quarantine" from March 1 to April 21 in 2020. We used a machine learning approach, Latent Dirichlet Allocation (LDA), to identify popular unigram, bigrams, salient topics and themes, and sentiments in the collected Tweets.

Results: Popular unigrams included "virus," "lockdown," and "quarantine." Popular bigrams included "COVID-19," "stay home," "corona virus," "social distancing," and "new cases." We identified 13 discussion topics and categorized them into different themes, such as "Measures to slow the spread of COVID-19," "Quarantine and shelter-in-place order in the U.S.," "COVID-19 in New York," "Virus misinformation and fake news," "A need for a vaccine to stop the spread," "Protest against the lockdown," and "Coronavirus new cases and deaths." The dominant sentiments for the spread of coronavirus were *anticipation* that measures that can be taken, followed by a mixed feeling of trust, anger, and fear for different topics. The public revealed a significant feeling of *fear* when they discussed the coronavirus new cases and deaths.

Conclusion: The study concludes that Twitter continues to be an essential source for infodemiology study by tracking rapidly evolving public sentiment and measuring public interests and concerns. Already emerged pandemic fear, stigma, and mental health concerns may continue

to influence public trust when there occurs a second wave of COVID-19 or a new surge of the imminent pandemic. Hearing and reacting to real concerns from the public can enhance trust between the healthcare systems and the public as well as prepare for a future public health emergency.

Introduction

More than four million people were confirmed positive of COVID-19 across 110 countries as of May 2020, and the death toll has reached close to 300,000 [1]. The widespread utilization of social media, such as Twitter, accelerates the process of exchanging information and expressing opinions about public events and health crises. COVID-19 is one of the trending topics on Twitter in the past four months since the outbreak. Since quarantine measures have been implemented across most countries (e.g., the Shelter-in-Place order in the United States), people have been increasingly relying on different social media platforms to receive news and express opinions. Twitter data are valuable in revealing public discussions and sentiments to interesting topics, and real-time news updates in global pandemics, such as H1N1 and Ebola [2-5]. In the current COVID-19 pandemic, many government officials worldwide are using Twitter, as one of the main communication channels, to regularly share policy updates and news related to COVID-19 to the general public [6].

Although there has been a growing body of empirical literature examining issues related to COVID-19 using Twitter data [6 7], most study samples were small, and therefore analyses using large samples of Tweets remain to be scant. Furthermore, methodologically, data processing and analysis were mainly relying on traditional qualitative coding techniques to make meaning of tweets. To extend the literature on public responses to COVID-19, the present study examines public responses in the face of the pandemic by analyzing approximate 22 million Tweets collected between March 1 and April 21, 2020. We integrated both unsupervised machine learning methods and qualitative coding techniques to triangulate the findings.

The present study aims to examine: (1) What the public discusses about the COVID-19 outbreak? (2) What are the communication patterns of the Twitter-based discussions about the pandemic? And (3) What are the concerns the public has expressed about the COVID-19 pandemic?

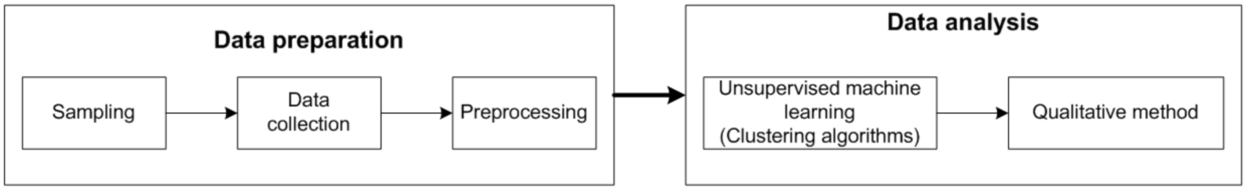
Methods

Research design

We used a purposive sampling approach to collect COVID-19 related Tweets published between March 1 and April 21, 2020. Our Twitter data mining approach followed the pipeline[1] displayed in Figure 1. Data preparation included three steps (1) sampling, (2) data collection, and (3) pre-processing the raw data. The data analysis stage included unsupervised machine learning, sentiment analysis and qualitative method. The unit of analysis was each message-level tweet. Unsupervised learning is one approach in machine learning, and used to examine data for patterns, and derives a probabilistic clustering based on the text data. We chose unsupervised learning because it is commonly used when existing studies have little observations or insights of the unstructured text data [8]. Since a qualitative approach has challenges analyzing large-scale Twitter data, unsupervised learning allows us to conduct exploratory analyses of large text data in social science research. In the present study, we first employed an unsupervised machine learning approach to identify salient latent topics. Using the topics, we used a qualitative approach to develop themes further, as a qualitative approach allows a deeper dive into the data, such as through manual coding and inductively developing themes based on the latent topics generated by machine learning algorithms.

[1] The pipeline was developed by the Artificial Intelligence for Social Justice Lab at the University of Toronto.

Figure 1. Twitter data mining pipeline



Sampling and Data collection

We used a list of COVID-19 related hashtags as search terms to fetch tweets, such as "#coronavirus," "#2019nCoV," "#COVID19," "#coronaoutbreak," and "#quarantine," from January 20 to present. Twitter's open application programming interface (API) allowed us to collect updated Twitter messages that are set open by default. From March 7 to April 21, 2020, we collected about 22 million (n=22,076,833) tweets during this period. After removing the duplicates and Retweets[2], we have approximately 4 million (n=4,196,020) Tweets in our final dataset. We collected and downloaded the following features for each tweet, including (1) tweet full text, (2) the numbers of favorites, followers, and friends, (3) user' geolocation; and (4) user' description/self-created profile.

Pre-processing the raw data

Shown in Figure 1, we used Python to clean the raw data, such as removing all non-English characters, the hashtag symbol, and its content, repeated words, and special characters, punctuations, and numbers from the dataset. More details were discussed in previous work [9].

[2] The Retweet message only reposts the original message without adding any more words.

Data analysis

Unsupervised machine learning

Latent Dirichlet Allocation (LDA) [10] is one of the widely used unsupervised machine learning approaches allowing researchers to analyze unstructured text data (e.g., Twitter messages). Based on the data itself, the algorithm produces frequently mentioned pairs of words, the pairs of words co-occur together, and the latent topics and their distributions over topics in the document [11]. Existing studies have indicated the feasibility of using LDA in identifying the patterns and themes of the Tweets texts related to COVID-19 [12 13].

Qualitative analysis

To triangulate and contextualize findings from the LDA model, we employed a qualitative approach to develop themes further. Specifically, using Braun and Clarke's [14] six steps of thematic analysis: (1) getting familiar with the keyword data, (2) generating initial codes, (3) searching for themes, (4) reviewing potential themes, (5) defining themes, and (6) reporting. Since the thematic approach relies on human interpretation, a process that can be significantly influenced by the personal understanding of the topics and a variety of bias, we have two team members conduct the first five steps independently. Then, the two members reviewed all independently identified themes and resolved disagreements together. Finally, we finalized themes corresponding to each one of the 13 topics.

Sentiment analysis

We used sentiment analysis, a natural language processing (NLP) approach, to classify the main sentiments of a given twitter message, such as fear and joy [15]. In the study, we used the NRC Emotion Lexicon, which consists of eight primary emotions: anger, anticipation, fear, surprise, sadness, joy, disgust, and trust [16]. We followed the four steps to calculate the emotion index for each twitter message, including (1) removing articles, pronouns (e.g., "and," "the," or "to"), (2) applying a stemmer by removing the predefined list of prefixes and suffixes (e.g., “running” after stemming becomes "run") [17], and (3) calculating the emotion index. We only keep one emotion with the maximum matching counts if one sentence has multiple emotions; and (4) calculating the scores for each eight-emotion type. We discussed the four steps in detail in the previous study [9].

Results

Descriptive results

A total of four million (n=4,196,020) Tweets consists of our final dataset after pre-processing all raw data (e.g., removing the duplicates). These Tweets were posted by Twitter users between March 7 and April 21, 2020. We identified the most popular tweeted bigrams (pairs of words) related to COVID-19. Bigrams captured “two concessive words regardless of the grammar structure and semantic meaning and may not be self-explanatory” [18], including “covid 19,” "stay home," "social distancing," "new cases," "don't know," "confirmed cases," "home order," "New York," "tested positive," "death toll," and "stay safe." Popular unigrams included virus, lockdown, quarantine, people, new, home, like, stay, don't, and cases. We presented the most popular unigrams and bigrams related to COVID-19 in Table 1 and visualized them using the word clouds in figure 2 and figure 3.

Table 1. Top 50 popular bigrams and unigram and their distributions

Top 50 bigrams	Dataset (%)	Top 50 unigrams	Dataset (%)
covid 19	0.29%	virus	1.18%
stay home	0.26%	lockdown	0.98%
corona virus	0.12%	quarantine	0.94%
social distancing	0.08%	people	0.82%
new cases	0.07%	coronavirus	0.79%
dont know	0.04%	new	0.47%
confirmed cases	0.04%	home	0.45%

home order	0.04%	like	0.44%
new york	0.04%	im	0.41%
tested positive	0.04%	stay	0.41%
death toll	0.04%	dont	0.41%
home orders	0.04%	cases	0.37%
quarantine got	0.03%	time	0.36%
stay safe	0.03%	covid	0.35%
spread virus	0.03%	19	0.30%
coronavirus cases	0.03%	need	0.30%
shelter place	0.03%	day	0.29%
coronavirus pandemic	0.03%	trump	0.28%
year old	0.03%	china	0.28%
public health	0.03%	know	0.28%
chinese virus	0.03%	going	0.25%
ill deliver	0.03%	help	0.25%
deliver copy	0.03%	pandemic	0.24%
health care	0.03%	world	0.24%
support usps	0.03%	health	0.23%
signing support	0.02%	think	0.22%
usps ill	0.02%	deaths	0.21%
wuhan virus	0.02%	today	0.21%
quarantine im	0.02%	good	0.20%
mental health	0.02%	work	0.20%
dont want	0.02%	want	0.19%
im going	0.02%	corona	0.17%
president trump	0.02%	spread	0.17%
united states	0.02%	got	0.17%
dont think	0.02%	support	0.17%
copy officials	0.02%	government	0.17%
feel like	0.02%	right	0.15%
looks like	0.02%	way	0.15%
positive cases	0.02%	care	0.15%
staying home	0.02%	social	0.15%
officials toodelivered	0.02%	news	0.15%
coronavirus outbreak	0.02%	state	0.15%
domestic violence	0.02%	country	0.15%
coronavirus lockdown	0.02%	said	0.14%
healthcare workers	0.02%	ive	0.14%
people died	0.02%	days	0.14%
quarantine day	0.02%	testing	0.14%
donald trump	0.02%	stop	0.13%
social media	0.02%	says	0.13%

Figure 2. The word cloud of most popular unigram

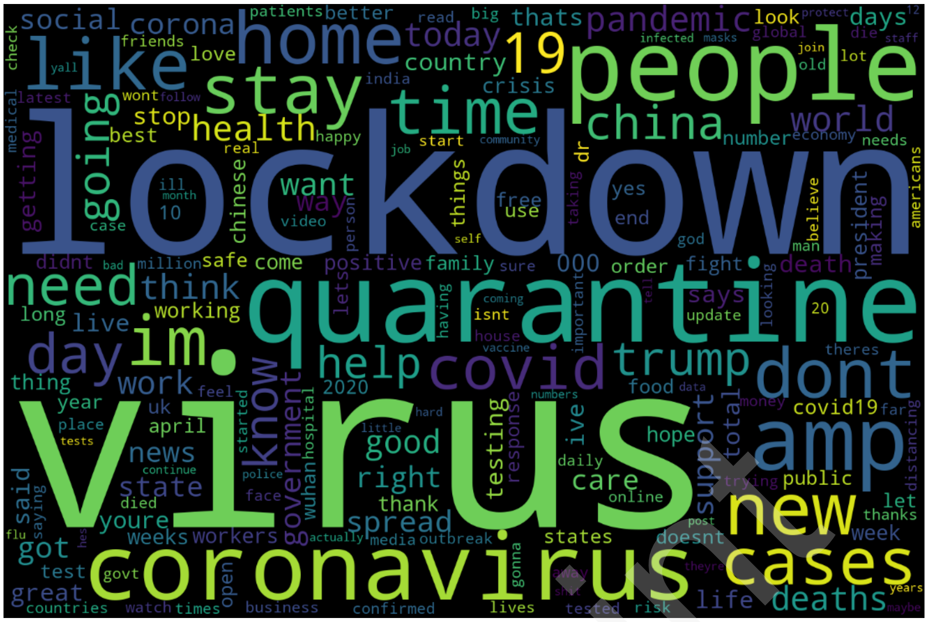
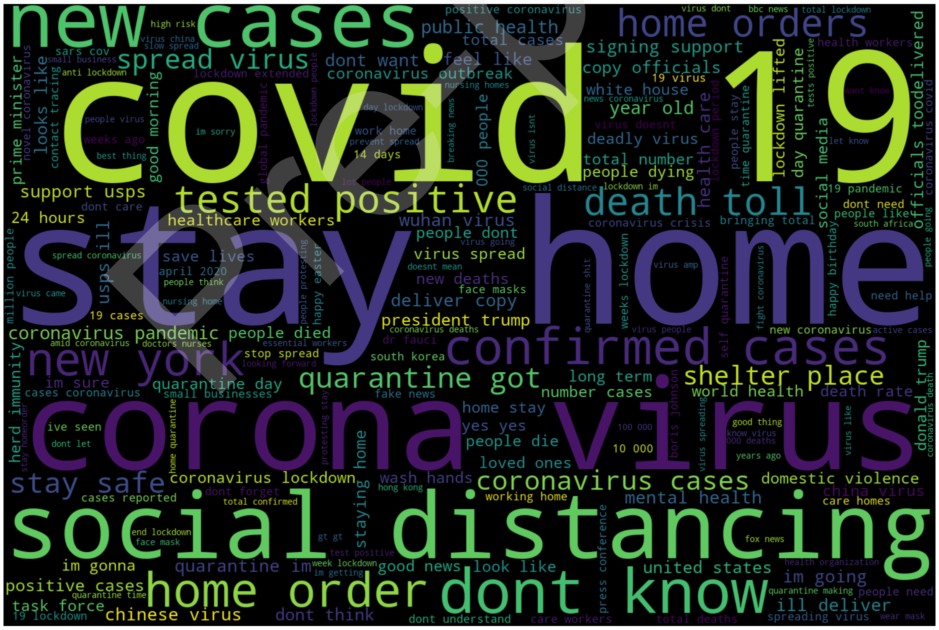


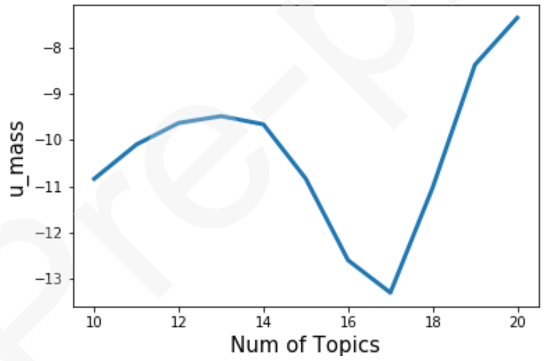
Figure 3. The word cloud of most popular bigrams



COVID-19 related topics

Our approach, Latent Dirichlet Allocation (LDA), produced frequently co-occurred pairs of words related to COVID-19. We organized these co-occurring words into different topics. LDA allowed researchers to manually define the number of topics (e.g., ten topics, twenty topics) that we would like to generate. Consistent with the previous studies, we used the coherence model – gensim [19] to calculate the most appropriate number of topics based on the specific data itself. For this dataset, the number of topics (n=13) returned by LDA had the highest coherence score as well as the smallest topic number. For example, the numbers of topics (n=19 or n=20) had higher coherence scores than the number of topics (n=13), but they represented larger topic numbers, shown in Figure 4.

Figure 4. The number of topics based on the coherence model



We further analyzed the document-term matrix and obtained the distributions of 13 topics with the chosen thirteen topics. We presented the results of the identified 13 salient topics and the most popular pairs of words within each topic in Table 2. For example, Topic 3 had the highest distribution (8.87%) among all 13 common latent topics. Within Topic 3, these pairs of words tend to co-occur together and share the same Topic 3, such as "tested positive," "coronavirus outbreak," "New York," "shelter place," and "mental health."

Table 2. Identified salient topics and their components (bi-grams)

Topic	Bigrams within topics	Distribution (%)
1	covid 19, dont know, deadly virus, im gonna, spreading virus, 19 lockdown, herd immunity, 000 people,19 pandemic, dont need, face masks, fox news, health workers, small businesses, home quarantine, like this, virus came, slow spread, test kits, total confirmed	8.51%
2	spread virus, health care, staying home, white house, positive cases, people die, 14 days, coronavirus deaths, care workers, ive seen, need help, day lockdown, know virus, im getting, doctors nurses, quarantine period, virus world, stop virus, people getting, week quarantine	7.24%
3	tested positive, coronavirus outbreak, wuhan virus, positive coronavirus, confirmed cases, new york, shelter place, mental health, china virus, feel like, new cases, gt gt, coronavirus covid, virus, weeks, people virus, people don't, bringing total, press conference, sars cov	8.87%
4	dont think, virus spread, lockdown period, fake news, nursing homes, wuhan lab, best thing, months, lockdown amp, 21 3, id like, people know, real time, entire world, know im, know it, wake up, feel free, dont wanna, anthony fauci	6.56%
5	u s, coronavirus cases, public health, save lives, novel coronavirus, long term, south korea, dont forget, bbc news, care homes, news coronavirus, million people, doesnt mean, family members, want know, coronavirus vaccine, going on, rest world, coronavirus, new jersey	7.36%
6	at home, stay at, home order, thank you, look like, good news, test positive, people stay, fight virus, people protesting, face mask, good thing, young people, lock down, wearing masks, cases deaths, trump said, deaths reported, shut down, active cases	7.36%
7	social distancing, day quarantine, healthcare workers, prime minister, world health, dont care, global pandemic, dont understand, health organization, dr fauci, let know, time lockdown, virus isn't, in place, anti lockdown, shelter in, people think, live updates, 2 months	7.81%
8	coronavirus lockdown, coronavirus crisis, amid coronavirus, looks like, new coronavirus, task force, im sure, coronavirus patients, prevent spread, virus doesn't, dont let, long time, new York, high risk, coronavirus task, thank god, number deaths, dont like, virus outbreak, coronavirus cases	7.47%

9	stay safe, chinese virus, self quarantine, need know, people going, new virus, common sense, safe stay, virus amp, b c, 2 2, family friends, we've got, got virus, stay away, testing kits, health amp, virus gone, april 20, knew virus	7.07%
10	corona virus, new cases, death toll, im going, quarantine day, people died, spread coronavirus, cases coronavirus, people dying, quarantine im, total number, number cases, cases reported, april 2020, confirmed cases, coronavirus death, 24 hours, people need, stop spread	**8.84%**
11	stay home, home orders, president trump, social media, home stay, loved ones, stay safe, death rate, working home, 31 000, social distance, 3100 000, protesting stay, breaking news, deaths, im sorry, 10 000, mortality rate	8.67%
12	coronavirus pandemic, year old, united states, wash hands, people like, work home, god bless, lot people, wear mask, years ago, virus hoax, like virus, 23 days, grocery store, said virus, 21 million, watch video, 10 days, like amp, uk lockdown	7.06%
13	right now, dont want, 3 weeks, tests positive, donald trump, weeks ago, weeks lockdown, virus spreading, coronavirus update, new zealand, 22 million, sounds like, total cases, lockdown 2, communist party, day day, chinese communist, cases 1, whats happening, 2 weeks	7.18%

Sentiment analysis of each latent topic

We presented the results of the sentiment analysis for each of the thirteen latent topics in Table 3. We also ran a one-tailed z test to examine if each of the eight emotions is statistically significantly different across topics. The *p*-value smaller than .01 was set as a threshold for significance.

Table 3. Percentage of each sentiment within 13 topics

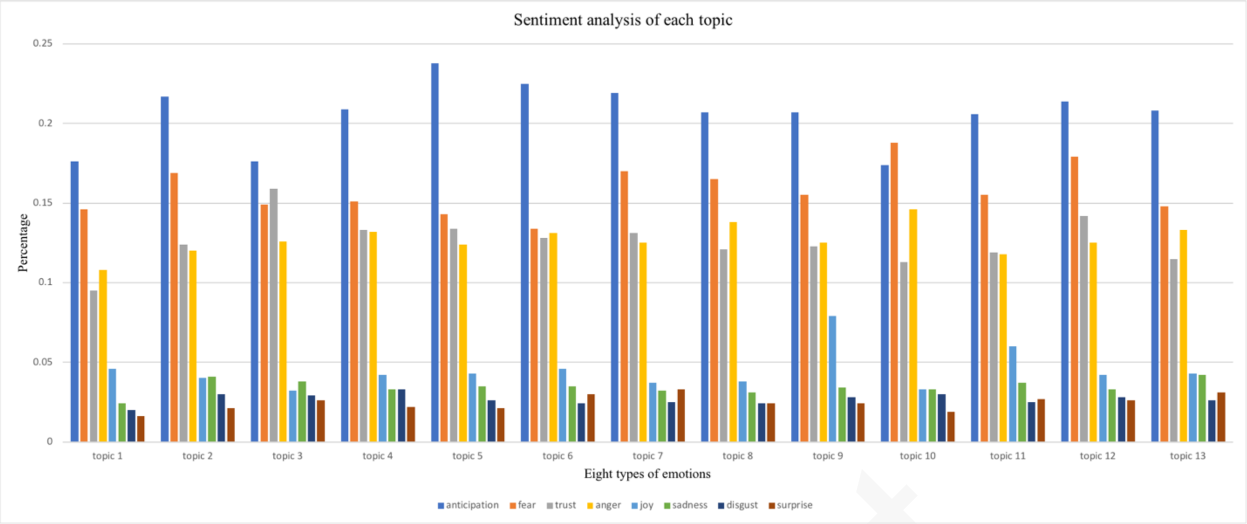
	Anger	Anticipation	Disgust	Fear	joy	sadness	Surprise	Trust
Topic 1	10.80%	**17.60%**	2.00%	**14.60%**	4.60%***	2.40%	1.60%	9.50%
Topic 2	12.00%	**21.70%*****	3.00%***	**16.90%*****	4.00%	4.10%***	2.10%	12.40%***
Topic 3	**12.60%*****	17.60%	2.90%***	14.90%	3.20%	3.80%***	2.60%***	**15.90%*****
Topic 4	13.20%***	**20.90%*****	3.30%***	15.10%	4.20%	3.30%***	2.20%***	**13.30%*****

Topic 5	12.40%***	**23.80%*** **	2.60%***	14.30%	4.30%	3.50%***	2.10%	**13.40%*** **
Topic 6	**13.10%*** **	**22.50%*** **	2.40%	13.40%	4.60%***	3.50%***	3.00%***	12.80%***
Topic 7	12.50%***	**21.90%*** **	2.50%***	**17.00%*** **	3.70%	3.20%***	3.30%***	13.10%***
Topic 8	13.80%***	**20.70%*** **	2.40%	**16.50%*** **	3.80%	3.10%***	2.40%***	12.10%***
Topic 9	**12.50%*** **	**20.70%*** **	2.80%***	15.50%	7.90%***	3.40%***	2.40%***	12.30%***
Topic 10	**14.60%*** **	17.40%	3.00%***	**18.80%*** **	3.30%	3.30%***	1.90%	11.30%
Topic 11	11.80%	**20.60%*** **	2.50%***	**15.50%*** **	6.00%***	3.70%***	2.70%***	11.90%***
Topic 12	12.50%***	**21.40%*** **	2.80%***	**17.90%*** **	4.20%	3.30%***	2.60%***	14.20%***
Topic 13	**13.30%*** **	**20.80%*** **	2.60%***	14.80%	4.30%	4.20%***	3.10%***	11.50%***

Notes: The sum of the percentage under each topic is not equal to 100%. The rests are either neutral or other emotions. *p*-value from Z-test, *** *p*< .001

Figure 5 presented eight emotions of trust, anticipation, joy, surprise, anger, fear, disgust, and sadness. Results showed that across all 13 topics, *anticipation* (dark blue line) dominated 12 topics, and followed by *fear* (orange line), *trust* (grey line), and *anger* (yellow line). For example, about 23.8% of Tweets in Topic 8 revealed feelings of anticipation that "necessary steps and precautions will be taken" [9 20]. The emotion *fear* for the impacts of the virus took 18.8% of the Tweets in Topic 10, which was statistically different from the fear expressed in other topics.

Figure 5. Sentiment analysis for each of the 13 latent topics

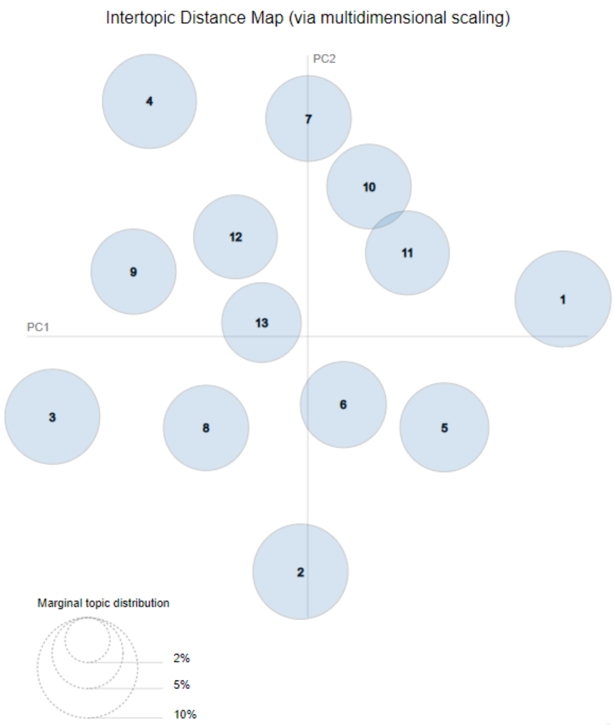


COVID-19 related themes

The qualitative content analysis approach allows users to categorize these topics into different distinct themes. Two team members discussed these bigrams and generated Tweets samples in each topic and then categorized the identified 13 topics into different themes. Besides, we computed the topic distance [10] to cross-validated the classification of the themes. Figure 6 showed a 2D plane of the intertopic distance[3] [21], in which each cycle represented a topic from Topic 1 to Topic 13 in the study. To protect the privacy and anonymity of the Twitter users, we did not present any user-related information, such as users' twitter handles or other identifying information; therefore, sample tweets shown in Table 4 were excerpts drawn from original tweets.

[3] . The circled areas are the overall prevalence, and the circle center was determined by computing the distance between topics.

Figure 6. Intertopic distance map[4]



Notes: PC: principal component
The principal component PC1 represents the transverse axix. The PC2 represents the longitudinal axis.

Table 4. Themes based on topic classification and sample tweets

Topic	Theme identified	Keywords	Sample tweets
1	Measures to slow the spread of COVID-19	covid19 lockdown, herd immunity, face masks, health workers, home quarantine, slow spread, test kits	"Testing Kits and USA: We urge the Modi govt to draw proper lessons from this latest instance of US…"
2	Quarantine and shelter-in place-order in the United States to stop the virus	staying home, white house, 14 days, day lockdown, quarantine period, stop virus, week quarantine, positive cases	"@realDonaldTrump @JustineTrudeau They'are all under mandatory 2 week quarantine, and they are essential workers… worry about boomers who can't follow instructions"
3	Mental health, and COVID-19 in New York	new york, shelter place, mental health, china virus, feel like, coronavirus covid, 19 virus, sars cov	"…New Yorkers on their apartment roofs during quarantine is a whole different vibe. This is gonna be in history books …"

4	Virus misinformation and fake news	dont think, virus spread, fake news, wuhan lab, best thing, people know, real time, entire world, know it, wake up, feel free, dont wanna, anthony fauci	"… Canada's COVID-19 research tied to Wuhan virus lab …" "…that China is responsible for putting entire world @great risk. Heavily criticized their eating habits."
5	Coronavirus cases in the rest of the world; a need for the vaccine to stop the spread	US, coronavirus cases, public health, save lives, long term, south korea, million people, going on, rest world, new jersey, coronavirus vaccine	"state of affairs in South Korea. Seen a lot of those posts saying capitalism so weak at the empty shelves after pani.. "…lead scientist for NIH working on #coronavirus vaccine research …"
6	Quarantine and staying home to stop the virus	at home, stay at, home order, thank you, look like, good news, test positive, people stay, fight virus, people protesting	"@funder: BREAKING: Every state without stay-at-home order had increase in coronavirus cases over the past week."
7	Protests against the lockdown	social distancing, day quarantine, healthcare workers, dont care, global pandemic, dont understand, health organization, dr fauci, time lockdown, virus isn't, in place, anti lockdown, shelter in, people think	"I stand with the Healthcare workers!!! Bravo! Healthcare workers face off against anti-lockdown protesters in Colorado …" "…nurses blocking anti lockdown protests in Denver …"
8	Task force in the US	task force, patients, prevent spread, virus doesn't, long time, high risk, coronavirus task, thank god, number deaths, dont like, virus outbreak	"RT @Jim_Jordan: There are #coronavirus task forces doing great work. But there is one task force that's missing in action: the U.S. congress.."
9	Got testing kits and stay safe and away from the virus	stay safe, self quarantine, need know, common sense, safe stay, family friends, we've got, got virus, stay away, testing kits, virus gone	"Testing Kits and USA: We urge the Modi govt to draw proper lessons from this latest instance of US..."

10	Coronavirus new cases and deaths	new cases, death toll, people died, spread coronavirus, cases coronavirus, people dying, total number, number cases, cases reported, confirmed cases, coronavirus death, people need,	"RT @neeratanden: 4,591 people died in a day from the virus, the highest number anywhere ever that we know of." "#Britain's death toll could be DOUBLE official tally as care homes"
11	President Trump's calling Chinese virus	stay home, home orders, president trump, social media, home stay, loved ones, stay safe, working home, social distance	"President Trump for failing to take action early on the #ChinaVirus Here are the facts …" "President Trump: They know where it came from. We all know where it came from, #chinesevirus..."
12	Coronavirus pandemic in the United States	united states, work home, god bless, lot people, wear mask, virus hoax, grocery store, 21 million	"stay-at-home orders continue in much of the United States …"
13	Chinese Communist Party (CPP) and the spread of the Virus	communist party, chinese communist, cases 1, whats happening	"Only way to Stop #ChinaVirus Made in #Wuhan is by Stopping the Chinese Communist Party = Stop Buying Made In China" "The #Chinese Communist Party (#CCP) is spreading disinformation to cover up the origin of the #coronavirus now infecti…"

A theme identified based on keywords in topic one was about public discussions around "measures or solutions to slow the spread of COVID-19." Keywords, such as "lockdown," "herd immunity," "face masks," "home quarantine," and "test kits," indicated this overarching theme. Topic 3 was about the public attention to the COVID-19 in New York and mental health concerns. We also identify other themes such as "misinformation and fake news (fake news, wuhan lab)," "need for vaccine (coronavirus vaccine)," "protests against the lockdown (healthcare workers, don't understand, anti-lockdown)," "coronavirus deaths (death toll, people die, people dying, coronavirus death)," and "coronavirus in the US (the United States, work home, god bless, wear mask, grocery store)."

Discussion

Principal results

The results show several essential points. First, the public is using a variety of terms referring to COVID-19, including "virus," "COVID 19," "coronavirus," "coronavirus." In addition, coronavirus has been referred to as the "China virus" that can create stigma and harm efforts to address the COVID-19 outbreak [22]. Second, discussions about the pandemic in New York are salient, and its associated public sentiments are *anger*. Third, public discussions about the Chinese Communist Party (PPC) and the spread of the virus emerged as new topics, which were not identified in a previous study using Twitter data collected between January 20 to March 7, 2020 [9], suggesting the connection between the COVID-19 and politics is increasingly to be circulating on Twitter as the situation evolves. Fourth, public sentiment on the spread of coronavirus was *anticipation* for the potential measures that can be taken and followed by a mixed feeling of trust, anger, and fear. Results suggest that the public was not surprised by the rapid spread of growth. Fifth, the public reveals a significant feeling of *fear* when they discuss the coronavirus crisis and deaths.

Comparison with prior work

Our findings are consistent with previous studies using social media data to assess the public health response and sentiments for COVID-19, and suggest that public attention has been focusing on the following topics since January 2020, including

(1) the confirmed cases and death rates [9 23];

(2) preventive measures [9 23];

(3) health authorities and government policies [6 9];

(4) daily life impacts such as food supplies and school closing [13 23];

(5) fake news and misinformation about the coronavirus [13];

(6) an outbreak in New York [9]; and

(7) COVID-19 stigma by referencing the coronavirus as the “Chinese virus” [22].

Compared with the study examining public discussions and concern for COVID-19 using Tweets from January 20 to March 7, 2020, we find that several salient topics are no longer popular, including (1) Outbreak in South Korea; (2) Diamond princess cruise; (3) economic impact [24]; and (4) supply chains [9]. Given the preventive measures, washing hands is no longer a prevalent topic. Instead, quarantine has become dominant.

In addition, our study identifies new discussion topics around the COVID-19 between March 7 to April 21, such as (1) a need for a vaccine to stop the spread; (2) quarantine and shelter-in-place order; (3) pretests against the lockdown; (4) coronavirus pandemic in the United States. The new salient topics suggest that Twitter users (English) are shifting their topics to more personally relevant or daily life-related conversations, rather than discussions about news content in South Korea, Diamond princess cruise, or Dr. Li Wenliang in China.

Limitations

First, we only sampled a trending of 25 hashtags as search terms to collect Twitter data. New hashtags keep coming up as the situation evolves. For example, a hashtag may become widely used only after a related topic becomes more popular, such as the official name, COVID-19, for the virus. Second, Twitter users are not representative of the whole population globally, and topics of Tweets only indicate online users' opinions about and reactions to COVID-19. However, the Twitter dataset is a valuable source allowing us to examine real-time Twitter user responses and online activities related to COVID-19. Third, non-English tweets were removed from our analyses, and hence the results are limited to users who posted in English only. Future studies are recommended to include other languages, such as Italian, French, Germany, and Spanish, for COVID-19 analyses.

Future research

First, future research could further explore public trust and confidence in existing measures and policies, which is essential. Compared to prior work, our study shows that Twitter users reveal a feeling of *joy* when talking about "herd immunity." We also find sentiments of *fear* and *anticipation* related to topics of quarantine and shelter-in-place. In addition, future studies could evaluate how government officials (e.g., President Trump) and international organizations (e.g., WHO) deliver and convey messages to the public, and its impact on the public opinions and sentiments. Finally, future studies could examine the spread of anti-Chinese/Asian sentiments social media and how people use social media platforms to resist and challenge COVID-19 stigma.

Conclusions

Studies have shown that Twitter data and machine learning approaches can be leveraged for assessing health communication research during the COVID-19 pandemic. Our findings facilitate an understanding of public discussions and concerns about COVID-19 pandemic among Twitter users between March 7 and April 21, 2020. *Anticipation* for the potential measures that can be taken to stop the spread of COVID-19 still significantly prevalent across all topics. However, Twitter users reveal *fear* when tweeting about COVID-19 new cases or death. As the situation evolves rapidly, new salient topics emerge accordingly. Real-time monitoring and assessment of the Twitter discussion and concerns can be promising for public health emergency responses and planning.

References

1. Center for Systems Science and Engineering (CSSE). COVIS-19 dashboard by CSSE at Johns Hopkins University (JHU). Accessed May 20, 2020: https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6
2. Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of tweets during the 2009 H1N1 Outbreak. PLoS One 2010;5(11):e14118.
3. Jones JH, Salathé M. Early Assessment of Anxiety and Behavioral Response to Novel Swine Origin Influenza A(H1N1). PLoS One 2009;4(12):e8032.
4. Kim Y, Kim JH. Using photos for public health communication: A computational analysis of the Centers for Disease Control and Prevention Instagram photos and public responses. Health Informatics Journal 2020:146045821989667.
5. Signorini A, Segre AM, Polgreen PM. The use of twitter to track levels of disease activity and public concern in the u.s. during the influenza h1n1 pandemic. PLoS one 2011;6(5):e19467.
6. Rufai SR, Bunce C. World leaders' usage of Twitter in response to the COVID-19 pandemic: a content analysis [published online ahead of print, 2020 Apr 20]. J Public Health (Oxf) 2020; fdaa049.
7. Kouzy R, Abi Jaoude J, Kraitem A, et al. Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on Twitter. Cureus 2020;12(3):e7255.
8. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. New York: Springer, 2013.

9. Xue J, Chen J, Chen C, Zheng CD, Zhu T. Michine learning on Big Data from Twitter to understand public reactions to COVID-19. arXiv 2020; 2005.08817.
10. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. Journal of Machine Learning Research March 2003;3:993-1022.
11. Xue J, Chen J, Gelles R. Using Data Mining Techniques to Examine Domestic Violence Topics on Twitter. Violence and Gender. 2019;6(2):105-114.
12. Cinelli M, Quattrociocchi W, Galeazzi A, et al. The covid-19 social media infodemic. arXiv 2020; 2003.05004.
13. Abd-Alrazaq A, Alhuwail D, Househ M, Hamdi M, Shah Z. Top concerns of Tweeters during the COVID-19 pandemic: infoveillance study. Journal of Medical Internet Research 2020;**22**(4):e19016.
14. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2006;**3**(2):77-101.
15. Beigi G, Hu X, Maciejewski R, Liu H. An overview of sentiment analysis in social media and its applications in disaster relief. In: Pedrycz W, Chen S-M, eds. Sentiment analysis and ontology engineering: an environment of computational intelligence. Cham: Springer International Publishing, 2016:313-40.
16. Mohammad SM, Turney PD. Nrc emotion lexicon. National Research Council, Canada. 2013.
17. Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. Cambridge: Cambridge University Press, 2008.
18. Xue J, Chen J, Gelles R. Using data mining techniques to examine domestic violence topics on twitter. Violence and Gender 2019;6(2):105-14.

19. Michael R, Andreas B, Alexander H. Exploring the space of topic coherence measures. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. Shanghai, China: Association for Computing Machinery, 2015:399–408.

20. Prabhakar Kaila D, Prasad DA. Informational flow on twitter–corona virus outbreak–topic modelling approach. International Journal of Advanced Research in Engineering and Technology (IJARET) 2020;11(3).

21. Chuang J, Ramage D, Manning C, Heer J. Interpretation and trust: designing model-driven visualizations for text analysis. Paper presented at: SIGCHI Conference on Human Factors in Computing Systems 2012; Austin, Texas.

22. 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. Journal of Medical Internet Research 2020;22(5):e19301.

23. Stokes DC, Andy A, Guntuku SC, Ungar LH, Merchant RM. Public priorities and concerns regarding COVID-19 in an online discussion forum: longitudinal topic modeling. Journal of general internal medicine 2020:1-4.

24. Medford RJ, Saleh SN, Sumarsono A, Perl TM, Lehmann CU. An" infodemic": leveraging high-volume Twitter data to understand public sentiment for the COVID-19 outbreak. medRxiv 2020.