Vaccine Hesitancy and COVID-19 Topics on Twitter during the Omicron outbreak

Stephanie Xie, Tianxi Zhuang, Lambert Francis
McGill University
845 Rue Sherbrooke O
Montreal, Quebec H3A 0G4

Introduction

Coronavirus disease 2019 (COVID-19) is caused by a novel and highly transmissible pathogen, named severe acute respiratory syndrome coronavirus type 2 (SARS-CoV-2). This deleterious disease was first officially reported in Wuhan city, Hubei province in China, on December 31, 2019. According to World Health Organization, as of December 7, 2021, there have been more than 265 million confirmed cases of COVID-19 worldwide, including more than 5 million deaths. Thankfully, in December 2019, several vaccines for COVID-19 had shown promising results in large trials and had been approved for emergency use. While we are drafting this report, more than 7.9 billion vaccine doses have been administered.

As people live through the COVID-19 pandemic, many issues have arisen with the virus rapidly spreading worldwide. These issues are related not only to the medical field, but also are tightly linked to the social, economic, environmental, and political domains. Although different types of vaccines have reduced viral load, a recent new COVID-19 variant has emerged in South Africa that appears vexatiously different from the one pathogen vaccines were designed to fight. As different strains of COVID-19 quickly spread over the world, discussions around the variants, vaccines, business, antigen testing, and government health measures are becoming the most talked about topics in various aspects of our daily lives. In particular, social media has rapidly become a crucial communication tool for this disease discussion.

One of the foremost popular online platforms for discussing those topics is Twitter. The tweets posted on Twitter are considered to be an excellent proxy for the public opinions and discernment associated with the widespread virus that the world is now encountering. Analyzing the tweets provides insights into understanding how people conceptualize and react towards this global crisis. In addition, investigating the Twitter discussions around the current pandemic is significant for understanding what aspects and concerns are perceived to be more salient and prominent for the general public. Therefore, this report aimed to utilize the individual tweets (not including retweets) to gain comprehensive insights on the salient topics discussed around COVID-19 and primarily concerns of each topic, the relative engagement with these topics, and the population sentiment polarity on these topics.

To begin with, one thousand original tweets are retrieved through Application Programming Interface (API), using the keyword “COVID19”, in English, from November 29, 2021, at 20:00 UTC to November 29, 2021, at 22:00 UTC. Through open coding on 200 tweets, we have identified eight topics. These topics are vaccine, COVID-19 variants (mutation), economic domain, political domain, health measures (response to pandemic), new cases reports, symptoms and antigen testing, and others. The tweet is marked as positive sentiment if it expresses an optimistic attitude during the pandemic and promotes vaccination. The tweet is annotated as negative sentiment if it complains or exhibits cynicism/skepticism towards each topic. Other tweets that do not fit either sentiment category are noted as neutral sentiments.

Through analyzing the annotations of one thousand tweets, we found that COVID-19 Twitter discussions consist of the following topics: vaccination, COVID-19 variant, economic recovery, political remarks, health measures, symptoms and antigen testing, and new case reports. By examining the number of posts related to each topic, we found that discourse around vaccination dominants over other domains. In particular, the number of negative sentiments towards vaccines are similar to those of the positive sentiments, which suggests that to promote vaccination among the general public, the not-for-profit organization should promote more scientific facts on social media.

Data

Through the publicly accessible Twitter Application Programming Interface (API) services, we are able to mine the tweets that global users post online, in compliance with the privacy regulations set by Twitter. The original data of 1,000 tweets were collected from Twitter using “COVID19” as the keyword in English from November 29, 2021, at 20:00 UTC to November 29, 2021, at 22:00 UTC. We could have used a keyword list including ”COVID19”, ”Corona”, ”SarsCoV2”, etc. to query tweets related to COVID-19. However, since we are only collecting 1,000 posts, using multiple keywords would shorten the time frame range of the collected posts, which may potentially lead to an inaccu-

rate representation of the results. To construct a balanced and representative collection of tweets, the tweets retrieved do not contain any retweets, and all tweets are unique through filtering tweet ID (same tweet IDs are excluded from the dataset). Although we have filtered repeated posts, we have noticed that some contents contained the same information yet with different users tagged, which escaped our filter. Nevertheless, this portion of the tweets only contributes to a small percentage of the total collected tweets; therefore, it would not jeopardize the quality analysis.

Table 1 demonstrates an example of the collected tweets with modification. Only tweets ID and contents (texts) are extracted from the JSON file as other information is non-essential for the annotation process. These values are constructed to a TSV file with the "Category" and "Sentiment" columns added for annotation purposes.

ID	Text	Category	Sentiment
14**12	Can anyone really believe these #COVID19 changes are natural viral evolutionary mutations	M	-1
14**32	Get your #COVID19 BOOSTERS! Push the boost.	V	1
14**95	What Happened: Dr. Jay Bhattacharya on 19 Months of #COVID19 https://t.co/kce1PWUkK2 via @YouTube	O	0

Table 1: Example of tweets collected with shortened tweets ID and contents (texts) pulled out. Category and Sentiment columns are for manual annotation purposes.
Category label: M: Mutation, V: Vaccine, O: Other
Sentiment label: 1: Positive, 0: Neutral, -1: Negative

Methods

Data Acquisition

Twitter has around 152 million active users worldwide and is a gold mine of data. By utilizing the API services, the Twitter platform allows analysts to retrieve the tweets that users post online within the privacy regulations set by the platform programmers. Notably, it allows users to do complex queries such as pulling every tweet about a certain topic within a certain period.

In order to achieve this, we simply need to build a query that satisfy our needs. The query we used, "%23COVID19 lang:en -is:retweet", searches for tweets that are tagged with COVID19 and the language is English and excludes retweets. The twitter fields we used, "tweet.fields=text,author_id,created_at,lang", simply filters out the information we need for our analysis. Then, we pulled the tweets from Twitter's API, with query, twitter_fields, start_time and end_time added to the url link. In order to check that every tweet collected is unique, we wrote a separate script to compare the tweet id using set.

In the beginning, we noticed that there were too many tweets with COVID19 keywords in the 3 days; thereby, we have only collected 1000 tweets from November 29, 2021, 20:00 UTC to 22:00 UTC. We chose the post period (created_at) from 20:00 UTC to 22:00 UTC because it converts to 15:00 EST to 17:00 EST and 12:00 PST to 14:00 PST, suggesting that this time frame is daytime for North America. In addition, this time frame corresponds to 9:00 NZDT to 11:00 NZDT, which implies that posts collected are in the morning of New Zealand. Furthermore, the corresponding time in Australia is from 7:00 to 9:00. This suggests that it is reasonable to use this specific time period to collect the posts because the five core Anglophone countries are in the daytime. More specifically, those times in each country are periods that Twitter has high amounts of engagement rate. Therefore, our tweets collected represent English-speaking countries' discussions around COVID-19.

Data Processing

Common english stopwords were excluded from the data as these words do not contribute useful information on the topics. The stopwords used can be found at `https://gist.githubusercontent.com/larsyencken/1440509/raw/53273c6c202b35ef00194d06751d8ef630e53df2/stopwords.txt`. In addition, we filtered words that contain non-alphabets such as "you're". With one exception, we chose to keep hashtag words, because there were some keywords such as "#Omicron". Then based on the word count results, we can easily get the tf (word, category), which stands for the number of the times that a word is mentioned under that category. The idf (word, script) is calculated by

$$idf_{(word,script)} = \frac{\log(total\ \#\ of\ categories)}{\#\ of\ categories\ mentioned}.$$

The tf-idf score is calculated using following equation:

$$tf-idf_{(word,\ category,\ script)} = tf_{(word,\ category)} \times idf_{(word,\ script)}$$

Ten words in each category with highest tf-idf scores are computed, but only the top three are listed here due to the space limit (See Table 2).

Data Annotation

At first, we conducted an open coding on the first 100 tweets. Two of us checked the content of each tweet and discussed what it is related to. Then we summarized the general topics based on the contents of the similar tweets. Later, we checked the correctness of the topics and made some adjustments based on another 100 tweets. For example, we combined the topic of COVID symptoms and testing to one topic and added a new topic of business. We ended up with 8 topics in total (See Table 2). In addition to topic categorizing,

Category	Most relevant	Second-most relevant	Third-most relevant
Politics	“political”	“#uspoli”	“#democrats”
Mutation	“surges”	“concern”	“dominant”
Other	“#orlando”	“#realestate”	“#florida”
Health	“staying”	“esstential”	“wearing”
Symptoms and testing	“test”	“detect”	“thermo”
Vaccination	“booster”	“transmission”	“mucosal”
New Cases	“usafacts”	“#datavisualization”	“#datascience”
Business	“zients”	“business”	“profitable”

Table 2: Top three words ranked by the value of tf-idf in each category. Ten words were listed originally in the data from high tf-idf to low tf-idf. Only three words are present due to space constraints.

we had also coded each post for positive/negative/neutral sentiments. One post was marked as positive sentiment if it promotes positive point of views during the pandemic or encourages government health measures. In regards to vaccination, positive sentiment posts were the ones that appealed to the public for getting vaccination/boosters. Negative sentiment posts criticized the government COVID-19 policies or showed cynicism towards their categorized topics. Tweets that do not fall in either sentiment group are marked as neutral sentiments. Notably, government tweet encouraging vaccinations are considered as neutral sentiments, as the target of interest is the general English-speaking public. After 1,000 tweets were annotated, we ensured the annotations were accurate and precise by double-checking each annotation individually. In some cases, we looked at the posts on Twitter to better categorize the topics and sentiments.

Results

Topic selection

The topics are defined as politics, mutation, health, symptoms and testing, vaccination, new cases, business/economic, and other. All represent the general idea of each tweet. For example, politics refers to tweets that criticize the reaction the government had to COVID19 or the new variant Omicron, or call for a new policy from the government. The tourism ban, mandatory vaccination and the WHO suggestions all relate to the political issues regarding COVID19. The mutation contains tweets with the variants of the COVID19, including Omicron and delta. It also could be the possible naming problems for future new mutated COVID19 viruses. The health stands for the safety measures that we should take during the pandemic, such as, wearing masks and staying at home. It could be a tweet that claims the benefit of working from home as well. The conjugated topic of symptoms and testing is related to the possible symptoms and the PCR test methods. We combined these two together because the testing and the symptoms are correlated in most cases. The possible tweet might be that I did a PCR test for COVID19 because I had fever. Other posts like supposing new rapid way of diagnosing COVID19 are categorized into this topic. The vaccination topic can vary from tweets to tweets. It can be ’I got the booster shot’ or any information regarding the vaccines. It also includes opinions on being vaccinated, positive or resistant. The new cases emphasize on the reported new confirmed cases in each city or country. It is more related with statistics, like how does the vaccination process affects the daily confirmed cases. The business topic is about the pharmaceutical company or the workers. Most tweets under this topic are related with the business plan of a company or unemployment of works under a company. The other topic includes all the tweets unrelated to COVID19 such as advertisements. The sentiment is coded as positive if the tweet expresses a positive attitude towards vaccination or the pandemic response. The tweet is annotated as negative sentiment if it criticizes the current policy or is anti-vaccine or shows a negative attitude towards the pandemic. Any tweets that don’t show a clear attitude are noted as neutral sentiments.

In Figure 1, almost four fifth of the data fell into one of topics, which means that the topics were well-chosen and covered most of the area of COVID19. Also the number of tweets under each topic is equally distributed except for the business and symptom and testing.

Topic characterization

Among the seven topics, other excluded, the vaccination has the largest number of tweets. This indicates that the vaccination problem is the hottest topic under the hashtag COVID19. The top 3 relevant words for the vaccination category are “booster”, “transmission”, and “mucosal”. It reflects that the public cares about the booster shot of the vaccine. And it might be how the booster shot can stop the virus transmission. Or it might be the influence of the mandatory policy of vaccination on the transmission of the COVID19. The word “mucosal” is a little bit mysterious. It could be one side effect of the vaccine.

The topics of new cases, politics, and mutations share similar distribution in the number of tweets. The confirmed cases are a big concern to the public. Since it is closely related to statistics, it is not surprising to see that the relevant words for that topic are ”usafacts”, “#datavisualization”, “#datascience”. It reflects that citizens in the USA are concerned about the daily new cases most, or the major portion of the users on Twitter during that pulling period resides in the US. This makes sense because the time we searched for the tweets is in the daytime in North America. The words that are most relevant to politics are very political, “political”, “#uspoli” and “#democrats”. This implies that people do have much to comment on the policy to COVID19, es-

pecially in the US. “surges”, “concern”, and “dominant” are the three words representative for the topic mutations. These words showed that people might be panic about the new variant Omicron as the number of confirmed cases surges. The results make sense because these three topics are also hot topics in real life, raising a lot of discussions on Twitter.

Health, business, and symptoms and testing are the three minor topics. One possible reason for this is that the pandemic has already lasted for two years and people are accustomed to the safety measures, PCR testing. It is expected that the top 3 words with high tf-idf values are “staying”, “essential” and “wearing”. People must be asking about the necessity of staying at home or wearing masks every day. For the symptoms and testing, the words, “test”, “detect”, “thermo” are also directly relevant to the topic. With “zients” listed as the most relevant word for business, it suggests that people have many comments regarding business during COVID19 to Jeffery Zients, who is the White House Coronavirus Coordinator. The word “profitable” indicates a positive attitude towards the business during the pandemic.

The other topic contains unrelated information like “orlando”, “realstate” and “florida”. It is most likely to be some advertisement on tourism or real estate.

Topic engagement

In general, topics regarding business, symptoms and testing, and vaccination have a similar amount of negative or positive sentiment tweets. But the amount of positive and negative tweets is significantly larger than any other topic. This implies that people tend to have a strong attitude towards vaccination. But the amount of people that support or resist vaccination is approximately similar. Vaccination is also proven to be the most controversial topic among all eight topics. The health has more positive tweets, showing that the public is willing to follow the current safety measure such as wearing masks and staying at home. The politics have even amounts of negative posts as vaccination regardless of its smaller total amount of tweets. This means that most of the public is complaining about the current policies, such as travel bans and mandatory vaccination. People are very unsatisfied with the government. It is very understandable because the pandemic lasts for two years and people might get tired of being locked down. The criticism might not just be to the government, but some national officials as well, like Trump. We think they might be anti-trump instead of solely blaming for the bad decision made for preventing the spread of COVID19.

The mutation and new cases also have more negative results than positive ones. The new variant Omicron, with a stronger ability to spread, would definitely cause some panic among the public. The result of increasing confirmed cases each day would no doubt amplify this panic. People might be worried about when they can take their masks off, can travel to different countries freely. Because the vaccination rate now is very high, but the number of confirmed cases is still rising every day, people must be worried about the efficacy of the vaccines.

In conclusion, the tweets are very representative of the public’s attitudes towards different aspects of COVID19. The fact that vaccination has the largest amount of tweets and is most controversial indicates that people care about vaccination most, from the booster shot to the efficacy. Also, it is important to mention that a significant amount of people complain about the policies from the government, which is largely caused by the long duration of the pandemic.

Discussion

As shown in Figure 1, the most discussed topics were vaccinations, mutations, politics, health, and business/economics, with vaccinations leading by a good margin. We will restrict ourselves primarily to these topics in the following discussion as we aim to analyze sentiment to the COVID-19 vaccination efforts and pandemic as a whole, rendering the symptoms (especially given its infrequent appearance) and new cases (Includes primarily data visualizations, neutral contents) categories less relevant.

First, we shall inspect the various topics of conversation within each category. Then we will analyze the sentiment of each category in the context of the internal topics. This approach, while hierarchical and simplistic, serves as a rough indication of the pandemic and vaccination response which is the objective of this report.

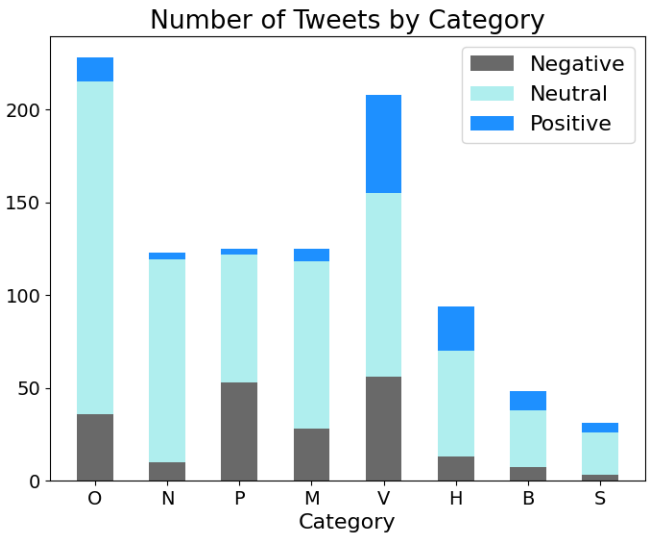

Figure 1: Bar chart depicting the number of tweets that belong to each category. Relative sentiment counts are also shown within each bar.

Vaccination

Vaccination being the most popular topic is unsurprising given its polarizing nature. Looking at the top tf-idf word in the vaccination category (See Table 2), we see the word ’booster’. This hints that the vaccination discussion primarily revolves around the recent rolling out of booster shots by Canadian and U.S. governments (whose citizens this analysis targets). Further down we see ’transmission’ and ’temperature’ which could be related to discussion of vaccinations with respect to the omicron variant (Temperature being

Category	Negative	Neutral	Positive	Percent Sentiment	Fraction Positive
O	36	179	13	0.27	0.27
N	10	109	4	0.13	0.29
P	53	69	3	0.81	0.05
M	28	90	7	0.39	0.2
V	56	99	53	1.10	0.49
H	13	57	24	0.65	0.65
B	7	31	10	0.55	0.59
S	3	23	5	0.35	0.62

Table 3: Tweet counts in each category by negative, neutral, and positive sentiment. Percent sentiment is the ratio of positive and negative tweets to neutral tweets. Fraction positive denotes the ratio of positive tweets to negative and positive tweets.

related to vaccine distribution in Africa, and transmission with curiosity about the transmissibility of the new variant).

Mutation

Similarly, the mutations category also generated heavy conversation likely due to the emerging omicron variant. Top words included 'surges', 'concern', 'dominant', and 'greek'. The first few indicate speculation about the transmissibility and effect of mutations on the pandemic as a whole. The appearance 'greek' can be attributed to the skipping of greek letters to omicron trending in social media.

Politics

Bipartisanship dominated the politics category with hashtags like '#democrats' and '#republicans' topping the tf-idf scores. This category was also a notable outlier with an overwhelming quantity negative tweets relative to positive tweets. This indicates that politics is a major driver in vaccine hesitancy as the negative positive coding focused on stance towards vaccination. Former President Trump appeared in three of the tf-idf words, suggesting that there still exist a lot of ongoing controversies among Trump's remarks about COVID-19.

Health

The health category words included 'staying', 'wearing', 'particles', and '#holiday'. We can infer from the middle two that conversation regarding masks was prevalent. The other two suggest that another relevant topic is vacation and travel in the holiday season. Considering that tweets in the health category were primarily of positive sentiment (The only category where this is the case), we can infer that mask wearing is being encouraged and people feel safe enough in the pandemic to consider traveling in the holidays.

Business

Discussion of the transition from Deborah Birx to Jeffrey Zients as the White House Coronavirus Coordinator was a major conversation topic with top tf-idf words 'zients' and 'birx'. Economic recovery was also a topic of interest. Interestingly, the sentiment of business tweets was largely positive. This could suggest that people are optimistic about the leadership and economic future in the context of the COVID-19 pandemic, however, given the varied topics it remains difficult to speculate what exactly was responsible for this positive outlook.

Response to Pandemic

Typically most tweets in each category were neutral. Looking at the ratio of positive and negative tweets to neutral tweets can serve as a proxy to the polarization of the topic. According to table 3, vaccination and politics dominated this area, with respective ratios of 1.10 and 0.81, with other categories being less than 0.5. This makes sense in the context of the bipartisanship we saw earlier in the politics section. The target of the negative sentiment is less clear in the vaccination sentiment, however likely relates to vaccine boosters.

Focusing more on whether topics were viewed in good light or not we can examine the ratio of positive tweets to positive and negative tweets. We find that the vaccination, health, new case, business, and symptom categories maintain high (Near 0.5 or greater) positive sentiment rates, while the politics and mutation categories are primarily negative.

To further investigate vaccination hesitancy, the number of positive and negative sentiments are more than similar, with 53 tweets and 56 tweets, respectively. These results suggest that there is still a large number of people that are skeptical about the vaccine. By examining the content of these tweets individually, we can see that many of them are promoted by bipartisan cooperation, failing to recognize the safety and effectiveness of the vaccine.

Conclusion

We found that the COVID-19 online discussion included primarily topics like vaccine boosters, omicron, concern about the pandemic worsening, masks, holidays and travel, administration, and economical recovery. While the overall response to vaccination efforts and the pandemic was leaning towards negative, much of this is likely driven by bipartisanship. A more detailed analysis of boosters and omicron could serve to better gauge vaccine hesitancy and general feelings towards the pandemic.

Improvements

Two major improvements that could be made to this analysis are a more sophisticated sampling of tweets and a more rigorous coding and annotating process.

Tweets were sampled by selecting tweets containing #COVID19. Given more time, it would be beneficial to understand how this might affect the sample. For instance, perhaps #COVID19 is a tweet used more often by democrats and #coronavirus by republicans, or is an unpopular term and #omicron would generate a more representative sample.

Creation of a design document and a longer coding process in place of a simple open coding on 200 tweets could greatly strengthen the verdicts an analysis of this type is able to give.

Additionally the sentiment coding could be greatly improved by separating it into two categories: One that focuses only on vaccination sentiment, and another on the pandemic sentiment. The mixing of the two in this analysis made inferences based on sentiment less tenable.

Furthermore, in order to obtain a more representative view of the global discussion surrounding COVID-19, more tweets should be collected at different points in time. Note that with more tweets collected, repetitious discussions would increase. Yet, one can simply eliminate it by only collecting one post from one individual Twitter user.

Group Member Contributions

Lambert was primarily responsible for the code, the discussion part of the writing, the graph, the category table, and some annotations. Stephanie and Zoe were primarily responsible for the annotations. Stephanie completed the introduction and the data sections of the report. Also, she modified some of the code, the annotations, and the some other parts of the writing. Zoe was responsible for the method and the results parts of the report and she had also taken care of the tf-idf table.