A Data Analytic Approach to COVID-19 Discussion and Sentiment on Twitter

Héctor Leos (260835543), Robert Dow (260786838), Solim LeGris (260807111)

McGill University

{hector.leosmendoza, robert.dow, solim.legris}@mail.mcgill.ca

Introduction

In this project we collected a sample of tweets with the goal of understanding the current conversations about the COVID-19 pandemic in the English-speaking world. This sample was collected using query keywords that are relevant to the pandemic and that could give us access to the different conversations around it, such as "covid", "lockdown", and "vaccine". We then conducted an open coding on a subset of our sample to gather the main topics on these discussions, striving to make our topics as well-defined as possible. These included topics such as "Politics and Global Affairs", "Science", "Pro-vaxx" and "Anti-vaxx".

Subsequently, we conducted a variety of data analyses to understand the type of engagement Twitter users have with these topics, as well as the sentiment with which they engage with them. We also performed TF-IDF analysis to obtain the ten most relevant words that characterize each topic and a network analysis to investigate the keywords that appear in the same tweets the most.

Our results show that COVID-related tweets generally have a negative sentiment. However, the group which seems to be the most negative of them all is the anti-vaxxers: almost 60% of their tweets are negative. We found that tweets about scientific research are the most neutral, and that pro-vaxx tweets occur equally in positive, neutral and negative tweets. Our TF-IDF analysis shows that tweets about scientific research and about politics mention Africa more often than any other place, probably due to the fact that the data was collected when Omicron was reported and investigated by African scientists. Tweets about health and safety measures mainly talk about lockdowns, while anti-vaxx discourse is best characterized by the use of the word "lied".

Furthermore, we analyzed concurrent usage of the keywords used to collect our data and found that "vaccine" and "covid" occur together most often relative to our other keywords. We discuss and link our different results to provide a synthesized overview of the overall response to vaccination and the pandemic as suggested by our data. Finally, we discuss some of the limitations of this data science project such as the technical limitations imposed by the basic Twitter API access we had for data collection.

Data

We sampled a total of 42,365 tweets using the Twitter API. We sent requests to the Twitter API for approximately 500 tweets every hour from 00:00 EST on November 27 to 23:59 EST on November 29, 2021 to obtain a good representation of COVID-19 related tweets throughout each day. The tweets were not sampled in real-time as preliminary sampling suggested that real-time sampling did not accurately represent engagement with the tweets with respect to likes, retweets and quotes. As a result, the latest tweet collected when our data collection script was run was at least twenty-four hours old. For the tweet requests, we used the following collection of keywords and filters to build our query: `covid`, `covid19`, `vaxx`, `vaccine`, `vaccination`, `vaccinated`, `omicron`, `antivaccine`, `pfizer`, `moderna`, `spikevax`, `biontech`, `janssen`, `johnson and johnson`, `lockdown`, `quarantine`, `pandemic`, `jab`, `vaxxed`, `-is:retweet` and `lang:en`. We determined this set of keywords by first generating a COVID-19 word bank and then selecting the most appropriate keywords given the theme and scope of our project. Given the constraints imposed by Twitter on non-academic developer accounts, we could not use geographical tags to collect tweets originating from a specific country, so instead we analyzed tweets from the English speaking world.

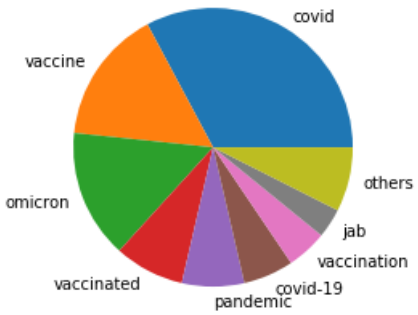

Figure 1: Keywords distribution within the tweet dataset. Each sector represents the relative frequency of each keyword in the tweets across the dataset.

During the exploratory phase, we determined the keywords with the most representation in our dataset, as shown in Figure 1. It should be noted that either "covid" or "omicron" occurred in half of the collected tweets – this is not surprising since the data was collected just a few days after the new variant was discovered. Other variants did not make it to the top eight most represented keywords. Moreover, a significant portion of tweets mentioned "vaccine", "vaccination" or "vaccinated", confirming that this was a suitable dataset for our purposes.

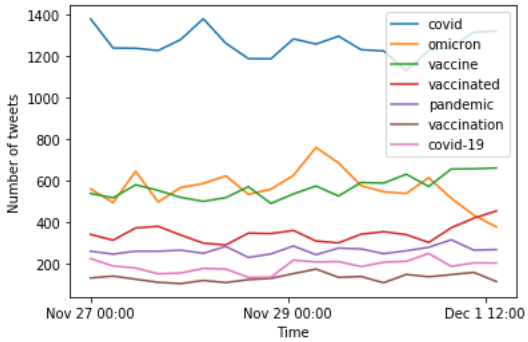

Figure 2: Keyword relative popularity over the data collection time frame.

Figure 2 depicts the progression of the collected tweets. At first glance, the conversations mentioning the Omicron variant seem to have the same relative popularity as those mentioning the vaccine with the trends following each other closely. Without access to data from before the emergence of Omicron, it's hard determine the nature of this relationship, so we decided to do some analysis to explore this relationship tangentially. From our collection of tweets, we sampled 1000 tweets uniformly without replacement. The resulting dataset was used in all our analyses. To simplify fetching tweets from the Twitter API, we used the Python library `tweepy`.

Methods

Before analyzing our data, we first cleaned all of the tweets by removing links and user mentions as well as emojis and ampersands which appeared as "&" in our data. Subsequently, words were lower-cased, stop words taken from (Landauer and Dumais 1997) were removed, and all punctuation except hashtags were removed.

Next, each team member separately conducted an open coding on 200 tweets from our dataset. We chose topics that would characterize differential views on vaccination, as well as those that would give insight into public opinion in topics of government, health and safety measures, and science. Together, these topics pertained to a majority of the tweets we collected, showing that they well characterize the types of conversations happening about COVID-19 on Twitter. A more detailed view of the topics used follows:

- **Measures** This included healthy and safety measures taken by the private and public sectors to control the spread of the virus. Tweets about event or program cancellations (due to COVID-19) were also considered in this category.
- **Politics and Global Affairs** People's opinion on the government's handling of the pandemic and mention of the global dynamics of this public health crisis.
- **Scientific research** This topic included any news or mentions about scientific breakthroughs on COVID-19, its variants and vaccine development, as well as any opinion from the lay public regarding this process in general.
- **Pro-vaxx** Opinions, comments or pieces of evidence shared by the general public in support of the vaccines against COVID-19.
- **Netural-vaxx** This topic includes tweets by the general public that mention vaccines but which don't take any particular stance.
- **Anti-vaxx** Opinions, comments or pieces of evidence shared by the general public against vaccines. Both conspiracy-like and science-based tweets were considered.
- **Other** All tweets not fitting in the previous categories. This included personal anecdotes, reports of COVID-19 cases, ads, etc.

After establishing these topics, all tweets were annotated as pertaining most strongly to one topic and a sentiment of negative, neutral, or positive was assigned to each tweet. For each topic, averages were calculated for each sentiment and net sentiment was calculated by (number of positive sentiment tweets) - (number of negative sentiment tweets) / (total tweets). To minimize subjectivity, we came up with rigid annotation guidelines. For instance, if anti-vaxx tweets were encouraging, we ascribed them a positive sentiment, even if we didn't agree with them. Same went with pro-vaxx tweets: they received a negative sentiment if for example they were being rude or discriminatory towards anti-vaxxers. Each team member annotated a third of the data, and we discussed ambiguous tweets among ourselves to secure a database with high validity.

These topics were then used in a TF-IDF analysis in which individual tweets were defined as documents. This analysis produced the ten most relevant and informative words that characterize each topic. Finally, the relative engagement with each tweet was evaluated by finding the mean number of likes, replies, retweets, and quotes of tweets in each topic. Note that for this step, we verified our dataset for outliers and found that one tweet had an abnormal amount of likes relative to other tweets, so we decided to remove it from this calculation.

Additional to the main analyses, and to better understand the relationship between tweets about Omicron and vaccination, we performed a network analysis with the Python library `networkx` using the original 42,365 tweets. The query keywords were used as nodes and the edges between them were determined to be the number of tweets in which both keywords appeared.

Results

Relative engagement

Table 1 displays the number of tweets occurring in each topic. Each topic contains between 63 and 305 tweets. Furthermore, these data serve as a metric of popularity of the topics in terms of what people are talking about on Twitter. The table also presents a more extensive analysis of the relative engagement in each topic. These results quantify the amount of interaction that is happening for each topic of discourse in the form of likes, retweets and quotes. Notably, these results indicate high engagement with tweets about politics and global affairs, as well as anti-vaxx tweets.

Topic	Count	Likes	Replies	Retweets	Quotes
Politics & Global Affairs	203	**11.81**	**2.13**	2.42	0.19
Measures	140	3.04	0.55	0.56	0.06
Science	101	4.04	1.52	1.2	0.17
Pro-vaxx	63	1.81	0.92	0.24	0.0
Ntrl-vaxx	82	2.16	0.56	0.37	0.04
Anti-vaxx	106	10.2	0.96	**11.03**	**1.29**
Other	305	3.34	0.52	0.55	0.1

Table 1: Number of topics and mean counts for each engagement metric.

TF-IDF analysis

Table 3 shows the results of the TF-IDF analysis which produced the ten most relevant words that characterize each topic. These words proved to be very informative in understanding the content discussed in each topic. For example, it was found that tweets on the topic of scientific research mentioned Africa and the chair of the South African Medical Association, Angelique Coetzee. This indicates that content about these keywords were particularly salient in this topic. This held true for measures, politics global affairs, and other. Notably, the three topics about vaccination are most informative in how they differentiate conversations about vaccinations. For example, pro-vaxx tweets talk about inconvenience and reduction, neutral-vaxx tweets mention proof and eradication, and protection, while anti-vaxx tweets mention lying and spread.

Sentiment analysis

Table 2 presents the findings of the sentiment analysis over all topics. These data indicate that pro-vaxx tweets and tweets about scientific research are balanced in terms of percentage positive and negative sentiment tweets. It also shows that a high proportion of scientific tweets are neutral in sentiment. Furthermore, anti-vaxx tweets are shown to be the most negative overall, with the least proportion of positive tweets. Lastly, it is shown that the tweets collected in our dataset are on average negative.

Topic	% Negative	% Neutral	% Positive	Net Sentiment
Politics & Global Affairs	45%	44%	10%	-35%
Measures	47%	39%	14%	-34%
Science	18%	62%	20%	2%
Pro-vaxx	32%	35%	33%	2%
Ntrl-vaxx	33%	55%	12%	-21%
Anti-vaxx	56%	36%	8%	-47%
Other	26%	59%	14%	-12%
Average	37%	47%	16%	-21%

Table 2: Average sentiment of tweets across topics.

Network analysis

The following show the top ten most connected pairs of nodes and their associated weights (representing the number of tweets in the original dataset in which the two given keywords appear):

- covid & vaccine: 813
- covid & omicron: 755
- covid-19 & omicron: 430
- covid & vaccinated: 409
- vaccine & covid-19: 357
- vaccine & vaccinated: 338
- vaccine & omicron: 280
- covid & pandemic: 215
- covid & vaccination: 178
- vaccine & pfizer: 166

The most popular word, unsurprisingly, was “covid”. The virus is mentioned alongside “vaccine”, “omicron”, “vaccinated”, “pandemic”, and “vaccination” in %17.3 of the total tweets which contain “covid” (13,720). This means that about a fifth of Twitter users in our dataset that mention the virus are also talking about its variants and about vaccines, which demonstrates how interconnected these topics are.

The next thing to notice is that “vaccine” and “omicron” appear in 280 tweets together. So the second thing people are mentioning the most when talking about vaccination (after covid) is the Omicron variant. This adds to the observation we previously made, when discussing Figure 2, regarding the strong relationship between conversations about vaccines and conversations about Omicron.

Discussion

As expected, the overall response to the pandemic and the topics we identified is relatively negative as indicated by our results (see Table 2). Our analysis suggests that the dominating topic in Twitter conversations surrounding the COVID-19 pandemic was politics and global affairs, representing 20% of tweets (see Table 1). More specifically, our data

Measures	Politics & Global Affairs	Science	Pro-vaxx	ntrl-vaxx	anti-vaxx	other
school	african	data	reduces	proof	lied	anagram
lockdown	biden	african	suck	eradicate	guaranteed	market
quarantine	meeting	coetzee	inconvenience	protected	spreading	folks
flight	countries	scientists	consistently	mania	data	yall
enforce	trump	study	faith	eradication	twice	smoking
procedures	britain	helpful	bigger	service	thousand	familiar
mandate	africa	chairwoman	vaxxed	pox	holy	moronic
schools	nations	current	survival	bn	stories	variation
staff	ban	click	trust	recommend	trials	na
travel	democracy	sequences	seriously	janssen	mutate	postpandemic

Table 3: Ten most relevant words for each topic, as calculated using TF-IDF.

suggests that the majority of the discussions for this topic are centered around information pertaining to the African continent (e.g. "african" and "africa"), American politics (e.g. "Biden" and "Trump") and international relations (e.g. "Britain", "nations", "countries"). These results mirror the current context involving the appearance and spread of the Omicron variant which was discovered in South Africa and certainly indicate a focus on this subtopic in Twitter conversations. Our data indicate that relative engagement was also highest for this category (see Table 1) with respect to likes (mean of 11.81) and replies (mean of 2.13) supporting the fact that a major portion of the COVID-related conversations on Twitter are politically or internationally oriented. Moreover, the overall response for this topic was negative with a net sentiment of -35%. Our data indicate that "measures" and "science" also occupied a major part of the conversation surrounding the pandemic with most of the discussion being centered around educational institutions, lockdowns/quarantine, vaccine mandates and travelling for the "measures" category. The net sentiment with respect to "measures" was negative with a value of -34% while interestingly, the net sentiment with respect to "science" was approximately neutral with a value of 2%. Relative engagement with these two topics was medium-low as suggested by (see Table 1).

The data suggests that the overall conversation surrounding vaccination is negative, especially when the stance or conversation is *against* vaccination, as indicated by our sentiment analysis results (see Table 2). Interestingly, tweets with a pro-vaxx stance are approximately neutral with a net sentiment of 2%. Although none of the three subtopics relating to vaccination that we chose for our design were by themselves the most popular topic, it should be noted that the majority of the tweets (251) come from these three categories. Consequently, a majority of the tweets surrounding the COVID-19 pandemic are relating to vaccination, but engagement with those tweets is relatively low except for "anti-vaxx" tweets. In fact, the highest mean retweet count was for tweets categorized as "anti-vaxx". The nature of this relative engagement (e.g. whether it is positive or negative) cannot be determined from our data and as such it would be interesting in further research to focus specifically on Twitter interactions involving anti-vaxx tweets. Our analysis shows that anti-vaxx tweets primarily involve distrust as the most popular word is "lied". Other popular words for anti-vaxx tweets such as "spreading", "data", "trials" and "mutate" suggest that the Twitter anti-vaxx conversation is also focused on the science of the vaccine and the virology of COVID-19.

We also gained some insight regarding the usage of our selected COVID-19 keywords in the different discourses surrounding the pandemic based on the conversations that are taking place on social media. It's clear that words such as "vaccine" and "pandemic" weren't as popular in everyday language usage a few years ago. Our analysis presents evidence for the concurrent use of certain words characteristic of the current context. For example, words such as "vaccine" and "omicron" were used extensively together. People are talking about the need to vaccinate a bigger portion of the global population to contain the spread of the Omicron variant and prevent other variants from arising. Conversely, other people are also pointing to the fact that the current vaccines may not be as effective against the new variant. It's interesting to note that, according to our TF-IDF analysis, people with a pro-vaxx stance use the word "consistently" a lot. We infer that this is probably the case when conversations take place about the constant effectiveness of vaccines against the virus and its mutations since mass vaccination began. On the other hand, anti-vaxxers seem to tend to put more emphasis on "mutate" indicating a potential lack of trust in the effectiveness of vaccines at stopping the ongoing pandemic. In contrast to the the often more extreme views of the pro-vaxx and anti-vaxx communities (e.g. overconfidence and distrust in vaccine effectiveness), people engaging in neutral discourse about the vaccine tend to have a more evidence-based perspective, as suggested by the popularity of the word "proof" for this category of tweets.

Although our results shed light on interesting insights surrounding the response to the pandemic and vaccination, our study was limited by a small dataset collected over a very short period of time. Consequently, it is unclear how sig-

nificant our results are in the broader COVID-19 context from the start of the pandemic to now and we do not have a good picture of how the response to vaccination and the pandemic may have changed over the course of the last two years. Moreover, we recognize that the overwhelming negative sentiment observed for anti-vaxx tweets may have partly been the result of the implicit bias in our team members' perception of anti-vaxx tweets as inherently expressing a negative sentiment. The data was collected at an unusual time since we have recently witnessed the appearance and spread of a novel COVID-19 strain which necessarily gave rise to uncertainty, fear, and doubt. This may have biased Twitter discussions more negatively than we would otherwise have observed throughout the pandemic. We also think that further network analysis of the top words by topic as determined by the TF-IDF analysis could have been insightful in understanding how and if these words co-occur. Lastly, due to technical limitations, our results are not specific to discussions around COVID-19 in Canadian social media but merely to the anglophone world. Clearly, our dataset was biased towards Omicron because it was extracted just a few days after the variant emerged. In further research (with access to a Twitter API for Academic Research), the reaction to the emergence of previous variants in social media should be investigated and compared to our current analysis.

Group member contributions

All team members contributed equally to the following: open coding, annotation and organizing the project. Solim wrote the code to collect the data from the Twitter API and wrote part of the data and the discussion sections of this document. Robbie wrote the code to clean and analyze the annotated dataset and wrote part of the methods and results sections of this document. Héctor performed exploratory data analysis as well as analysis with respect to relative engagement and wrote the introduction and parts of the data and discussion sections of this document.

References

Landauer, T. K.; and Dumais, S. T. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2): 211.