Comp598 Final Report

Claris Gu, Alexander Bertrand, Sabrina Yan
McGill University
jia.gu@mail.mcgill.ca, alexander.bertrand@mail.mcgill.ca, haihui.yan@mail.mcgill.ca

Introduction

Coronavirus has been identified as a deadly pathogen for both human and animal. In February 2020, the World Health Organization designated the disease COVID-19, which stands for coronavirus disease 2019. The virus that causes COVID-19 is designated to be virus that cause a severe acute respiratory syndrome and it was named as coronavirus 2 as well as SARS-CoV-2. Until now, it has been out for 2 years and during this period, there have been variants. A variant is a slightly altered - or mutated - version of a virus. There are thousands of Covid variants around the world, which is to be expected because viruses mutate all the time. The latest variant discovered is the Omicron, which was first discovered only a few months ago. It is strikingly different from many other types due to the long list of genetic mutations it has undergone. It is critical to access to safe and effective vaccines in order to end the COVID-19 pandemic, so there are many vaccines proving and going into development. Safe and effective vaccines are a game-changing tool, as well as follow the protection rule such as wear masks, clean hands, physically distancing and good ventilation indoors. However, a major roadblock to widespread vaccine adoption is vaccine hesitancy that spreads through social media. Our team would like to analyze the current situation of COVID in Canada by collecting tweets within a 3-day windows from Twitter. Here, the major goal is to understand the discussions currently happening around COVID in Canadian social media twitter. Specifically, we want to know the following: 1) the salient topics discussed around COVID/pandemic and what each topic primarily concerns; 2) relative engagement with those topics 3) how positive/negative the response to the pandemic/vaccination has been.

In this report, we aimed to investigate the most salient topics discussed around COVID-19 and more specifically, we were to find out what had been the reason behind vaccination hesitancy in Canada. To achieve this goal, 1,000 tweets were live streamed and filtered to restrict to Canada-originated tweets. These filtered tweets were then manually labeled with relevant categories and sentiments. Two most discussed topics were determined to be covid/pandemic and vaccine. These data were then used to get the top 10 words with the highest tf-idf score for each sentiment in each category. Then a set of selected words from the top words were used to search tweets in which the selected top word was found. This approach gave us insight into the reason behind people's reactions towards the two main categories: covid/pandemic and vaccine. The positive reactions from the covid/pandemic category mainly stemmed from fear of the death or damage that could be caused by the COVID-19 disease, and negative reactions were mainly stemmed from less fear or less awareness to the damages that could be cause by the COVID-19 disease. Positive reactions in vaccination mainly resulted from people's trust towards vaccine quality as well as fear of the disease, whereas negative reactions were mainly a result of fear and distrust to the quality of the vaccines.

Data

Tweets were live streamed from filtered stream endpoint using Twitter API v2 together with requests package written for Python. Tweets were first live-streamed through filtering tweets that including one of the keywords “covid”, “covid-19”, “vaccination”, “vaccine”, “Pfizer”, “Moderna”, “Johnson & Johnson vaccine” and one of “Canada” and “Canadian”. Additional rules on tweets to be collected include:

- Excluding retweets. Retweets were excluded to avoid repetitive tweets being collected.
- Some keywords were not included in the collected tweets: “Trump”, “Globalnews”, “CBC”. These keywords were excluded to avoid extremely-politics-biased tweets.
- tweets should not include hyper-links
- language was restricted to be English

First round of data collection yielded 1005 tweets. This round the data collection was then performed for another two days, each lasted for 20 to 24 hours using the same searching rules described above.

The second and third round of data collection yielded alltogether 1000 tweets, so the total number of data collected from filtered stream endpoint were 1005+1000=2005 tweets. All tweets were stored in json format with the following fields: “text”, which included the main text of the tweets, “user-location”, which includes all user-defined locations that can be accessed through a “user” parameter in the requests sent to the filtered stream endpoint.

All tweets were further filtered to restrict Canadian-only tweets. This second filtering step was done through iteratively inspect user-locations of each tweet stored as json object for Canadian provinces or major city names or their abbreviations. These included: “Canada”, “Ottawa”, “Alberta”, “Edmonton”, “Calgary”, “British Columbia”, “Victoria”, “Vancouver”, “Manitoba”, “Winnipeg”, “New Brunswick”, “Fredericton”, “Newfoundland”, “Labrador”, “St. John's”, “Nova Scotia”, “Halifax”, “Ontario”, “Toronto”, “Hamilton”, “Windsor”, “Mississauga”, “Kingston”, “Kitchener”, “Sudbury”, “Waterloo”, “waterloo”, “Prince Edward Island”, “Charlottetown”, “Québec”, “Québec City”, “Montréal”, “Saskatchewan”, “Regina”, “Saskatoon”, “Northwest Territories”, “Yellowknife”, “Nunavut”, “Iqaluit”, “Yukon”, “Whitehorse”, CA”, “AB”, “BC”, “MB”, “NB”, 'NL', “NS”, 'ON', 'PE', 'QC', SK”, “YT”, “NU”, “NT”. This list of names was obviously not extensive, but these names were selected after observing the input first-hand data for a better idea on which names should be used, and which ones never appear in the input data. The output of this second filtering step was stored as comma separated file, “complete_filter_data.csv”, which was later annotated in data annotation step. The final output after the filtering step yielded 1000 tweets.

Methods

We approached the initial open coding with a few main topic ideas in mind: we assumed there would be many discussions on vaccination, government, new variants, and mental health during the pandemic. The initial categorization was based on these codings, but we soon realized that there were more topics that needed to be added; many tweets we saw during the open discussed finances and business during the pandemic, so an “economy” category was added. Also, we had to remove some categories that weren’t getting many tweets, like the mental health category.

After a few rounds of open-coding, we eventually decided to investigate seven main topics: “vaccine”, relating to vaccines and vaccine passports; “economy”, for discussing the economic impact of COVID; “variant”, for tweets about new variants of coronavirus; “covid/pandemic”, for tweets regarding the disease’s impact on daily life; “movement”, for discussing limitations imposed on travel; “government”, for virus-related tweets directed specifically towards the government; and “precaution”, when talking about safety measures in place to limit coronavirus spread. We also used an “others” category for any tweet that did not fit into one of the other categories (these tweets were not used for analysis).

These labels were broad enough to fit nearly all the tweets in our dataset yet were each distinct enough that most tweets fell neatly into one category. In the handful of borderline cases that did occur, the final decision was entrusted to the coder.

As we continued into the data analysis phase, we noticed that when calculating our tf-idfs there were some common issues that arose in the outputs. By far the most common problem was that Twitter usernames were being counted as words due to @ mentions. While investigating which users are participating most in each topic may be interesting, it was outside the scope of this project, so words beginning with an “@” were excluded from the tf-idf calculation. We also chose to remove the hashtags from the dataset, since they are generally not part of the main content within a tweet. Finally, we removed the search keywords (described in “Data”) to avoid false-positives in the tf-idfs (since their abundance was artificially increased by us selecting for them).

Another challenge we faced was to come up with a satisfactory definition for the sentiment labels. Initially, we planned to simply score a tweet’s sentiment based on the surface-level “feeling” it gave; tweets that suggested the author felt “happy”, “excited”, etc. would be coded as positive sentiment. Tweets that indicated the author was “frustrated”, “angry”, etc. would be coded as having a negative sentiment. However, this resulted in us losing crucial context for many of the tweets.

For example, consider the tweets “Just convinced my brother not to get the jab!” and “My youngest just became eligible for the vaccine – we’re bringing him into the clinic tomorrow!”. On the surface, both sound like statements with positive sentiment due to their excited tone. Upon deeper inspection, though, the writers clearly hold opposite positions on the vaccine issue; we thought that it would be strange to classify contrary viewpoints in the same category.

Instead, we settled on a more objective definition of sentiment. Tweets would be labelled as positive if they “agree with the scientific community and/or the Canadian government’s pandemic response”. Tweets that opposed this viewpoint were labelled as negative, and tweets that did not have a clear opinion were labelled as neutral.

Results

The main topics used to label the tweets collected were covid/pandemic, government, movement, economy, vaccine, variant, precaution, others. Their corresponding definitions were listed as follows:

- covid/pandemic: for tweets relating to how coronavirus has impacted daily life.

- government: virus-related tweets directed at the government or members of the government about their coronavirus response.
- movement: for tweets relating to closing international/provincial borders and strict requirements for travel (such as vaccination and negative COVID test).
- economy: discussions about how COVID has affected the national or global economy.
- vaccine: for tweets relating to either the vaccines itself, or vaccine passports and other restrictions for unvaccinated individuals.
- variant: for discussions about new coronavirus variants and the dangers they may pose.
- precaution: for suggestions and discussions regarding safety measures to reduce the spread of COVID.
- others: for any tweet that does not belong in one of the above categories.

The relative engagement in each topic was calculated as the percentages of a label in all labels. The relative engagement for each category was listed in Table I.

Table I. category frequency and sentiment rates for each relevant category

category	frequency (%)	Positive rate(%)	negative rate(%)	neutral rate(%)
vaccine	22.57	15.33	40.67	44
covid/pand-emic	15.18	10.34	41.38	48.28
movement	8.30	15.56	62.22	22.22
economy	4.55	15.85	29.27	54.88
vaccine	3.44	28.12	37.5	34.38
variant	3.24	29.6	34.98	35.43
precaution	2.94	29.41	35.29	35.29
others	39.78	NA	NA	NA

NA was to indicate sentiment rates were not calculated for “other” category as it was not useful for sentiment analysis.

We found that the topic with the most engagement was the “vaccine” category, with 22.57% of all tweets belonging to it. Discussion on COVID and the pandemic in general was also relatively common, at 15.18%. The rest of the categories had significantly less activity; the most notable was “movement”, with 8.3% engagement, while others were all below 5%.

Positive reactions were defined as having an optimistic view on a particular category. For instance, a positive reaction in the vaccine category would be pro-vaccine comments. A negative reaction would then be having a pessimistic view on a particular topic. A negative reaction in the vaccine category would be con-vaccine comment.

Two major topics are covid/pandemic and vaccine based on their high relative engagement rates. For covid/pandemic category, 15.33% reactions were positive and 40.67% were negative. The discussion on vaccines was more balanced, with 29.6% reactions being positive and 34.98% being negative.

While the other topics were much less prevalent in the overall discussion surrounding COVID, some interesting results were seen. The sentiment towards the government was significantly more negative than for any other topic.

To investigate deeper in reasons for people’s reactions in each category, tf-idf score were calculated for each word in each topic for each sentiment category. The top 10 words with the highest tf-idf score were then used to characterize each sentiment category. The results were shown below for categories covid/pandemic and vaccine.

The top 10 words with the highest tf-idf score for each sentiment category were:

- positive: "approved", "toxic", "surface", "nanocyn", "kills", "fda", "epa", "disinfectant", "boeing", "approve".
- negative: "premature", "form", "nope", "death", "money", "heart", "fear", "disease", "left", "family".

“Approved” being the word that appear the most frequent in the positive category implied that people’s reaction towards

For the vaccine category, the top 10 words with the highest tf-idf score for each sentiment category:

- positive: "understand", "help", "doing", "approve", "answer", "anti", "quality", "mental", "mask", "living".
- negative: "immunity", "friends", "month", "vs", "unless", "trust", "reason", "lost", "hold", "freedom".

5 words among the highest top 10 words were selected to be used to search for tweets in the original collected dataset to gain more insights into the reason behind vaccine receptiveness or vaccine hesitancy. For the covid/pandemic category, the 5 words chosen for the positive category were “mrna", "approved", "weeks", "variant", "prevent”. The 5 words chosen for the negative category were "premature", "form", "death", "money", "heart". For the vaccine category, the 5 five words chose for the positive category were “help”, “approve”, “quality”, “mental” and “living”, and “immunity”, “friends”, “trust”, “reason” and “lost” were chosen for the negative category.

Using the 5 words selected for positive covid/pandemic-related tweets, the following tweets were found:

“A lot of people can't have the mRNAs and need a non-mRNA alternate. Even AZ and JandJ are mRNA. @GovCanHealth has a responsibility towards all Canadians.” This tweet showed support in testing for COVID-19 and was therefore labeled as positive.

Found tweets were also showing support for new vaccine for children such that “Yes where are those kids vaccines. That the US, Canada, Israel and the EU regulators have already approved. Has the TGA gone on holiday or something. Not like 1000s kids getting Covid in Vic and NSW is a problem or anything.” People’s positive reactions also showed in fear of being unvaccinated, such that “it's mostly the unvaxed dying in the hospitals.”

Using the 5 words selected for negative covid/pandemic-related tweets, the following tweets were found:

"…Should Canada worry about the 50% chance of premature death form cancer and heart disease before the 0.057% chance of premature death form covid? Gee where should the money go?" The write here compared the premature death from cancer VS. heart disease VS. covid. The data shown here is not official and certainly cannot be compared. As covid death is as important as other diseases' death. As a result, this would be labeled as negative.

"...a skeptic. But it's what so many people don't want to hear. Especially those who don't want to take responsibility for their health. Look at an excess death chart from Canada compared to other years. It's negligible. But live in fear if you choose. Seems like a good crutch." Here, the writer's attitude towards the death chart is indifferent as he/she stated as "negligible". The overall attitude of its statement is more toward "negative" response.

Discussion

By looking into the original tweets from which the selected words for each category came, insights could be gained for how people reacted to each topic.

Main results for insights gained for positive covid/pandemic reactions by searching top words with in the data were support for covid testing and call for new effective vaccines. People's support for overcoming the covid pandemic also showed in were also found to be the fear to be hospitalized or death caused by the virus. This is plausible as people who fear of the damages that could potentially be caused by COVID-19 viruses would be proactive in overcoming the pandemic by supporting vaccinations, testing or raising awareness online.

Negative covid/pandemic reactions were mainly a result of people awareness to the death or damage caused by the covid-19 virus. People who fear less of the virus and the disease had showed negative response to the covid pandemic topics. This is plausible since people will not be active in helping with overcoming the pandemic if they feel the virus was not a threat to their health.

Of the topics we investigated, vaccines were the most widely discussed. Interestingly, the proportion of positive and negative sentiments were very similar, which suggests that the topic is very hotly debated. This isn't too surprising, given that vaccines have been so widely politicized in recent years. Of users who do not support the vaccine, the most relevant word was "immunity". A common thread in arguments by vaccine detractors is that the herd immunity doctors have been promising if vaccinations were widely adopted hasn't come yet. Many people also hope the resistance to Coronavirus from overcoming an infection will help the population achieve herd immunity. As one user puts it, "...every human will contract Covid in one variant or another. Natural immunity will follow." However, this is a dangerous sentiment to have, as many hospitals and other healthcare facilities are already significantly over capacity.

Doubts of the vaccine's efficacy and fear of potential side-effects are also widespread. It seems the basis of this fear is often anecdotal evidence they hear from people close to them. "Friends" was the second most relevant word, and tweets alleging dangerous effects the vaccine has had on friends and relatives was common. For example, one user claimed "Vax killed one of my friends. My Canadian friend got the shot to come to US. But has health issue now." Whether or not these events are grounded in truth, it causes many people to have second thoughts about vaccination. Searching for the word "trust" also yielded some alarming tweets, including this one from a very suspicious Twitter user: "Not sure who to trust. Made the mistake of "trusting the science" and have been suffering since…Covid would have been much more manageable than whatever is in the vaccine being forced on us…". This type of response to the vaccine was extremely common.

Despite this negativity, there was also a significant amount of pro-vaccine tweets. Some users were hopeful that an increased vaccine rollout would not just prevent people from getting sick but also help stifle the virus' evolutionary progress, saying "…equity in distribution and to help stop the development of mutations…".

Additionally, there was some interest in a new Canadian-manufactured vaccine that will soon be released. One user stated, "…already producing novavax. There are several facilities around the world that are/will be producing it, Canada's facility is only one of them.". While other ones showed trust in the quality of the vaccine made in Canada, such as "…maintaining the same high standards for safety, efficacy and quality before approving." One more reason for being pro-vaccine was found to be a result of government's incentives for encouraging people to get vaccinated: "…shout out for giving salary increase to those who are vaccinated!"

It's also interesting that the government was viewed overwhelmingly negatively, with 62% of all posts having a negative sentiment and only 16% positive. The biggest factor that appeared to cause this was a lack of trust in the government, with "trust" being the most common word in negative tweets. That being said, there was some support among users for government policies; "vaccines", "flights", and "borders" were commonly used by those with a positive view of the government, which likely indicates support for the government's caution towards opening international borders as the Omicron variant spreads.

Overall, the data and the analysis are quite intriguing and could be used for further analysis in the future. It would be interesting to see how the same dataset could be used under different conditions.

Group member contribution

All members contributed to data annotation. Alex worked on the methods, discussion general editing of the report and wrote code for data cleaning and calculation of the tf-idfs. Claris worked on data collection, data cleaning, calculation of the tf-idfs. Sabrina worked on the introduction and some other parts of the report. We thank everyone for the amazing hard work and Professor Ruths for making the project which we really enjoyed it. We appreciate it!