# COMP598: Intro to Data Science
# Analysis of COVID-19 on Twitter

**Yifang Yuan, Muzhi Qi, Arman Izadi**
McGill University
3630 rue University
Montreal, Quebec H3A 0E9

## Introduction

This project aims to analyze society's sentiment and concerns towards the Covid-19 pandemic and its corresponding vaccine through analyzing Tweets from the Twitter social media platform. The gathered tweets were annotated manually into eight different topics: restrictions, demography, impact (impact on human life), vaccine, variant, symptoms, testing, and other. In addition, the sentiment of each tweet was taken into account by denoting them to be either positive, neutral or negative in relation to the defined topic.

Then, a TF-IDF analysis was ran to yield the top ten most important used words in each topical category. Finally, it is concluded that based on the gathered data and the analysis of sentiments as well as most important used words, the overall sentiment to the pandemic and the vaccines were that of a negative one. However, discussion related to the new Omicron variant seem to be very minimal, as the topic of variant isn't even one of the most engaged topic after comparing with other annotations. Also, the sentiment of the users regarding the vaccine fall on both spectrum. there are 59 of the 161 tweets related to vaccine that seem to be supportive, and only 83 that are negative.

Although the negative response rate is still higher, if we compare with the total positive - negative response rate ratio, this stands at 226 positive / 539 negative = 42 percent. The number of positive to negative ratio for the vaccine category stands at 59 / 83 = 71 percent, which is far more optimistic than the general negative trend. There were also a lot of political and social issues that were brought up by the users and most of them are negative.

## Data

During initial data collection, 1500 tweets were gathered via the Twitter API's search endpoint. Then, the tweets were reduced down to approximately 990 tweets by ignoring duplicates, and spam tweets from the same user in a very short time span. The keywords that were used to collect the tweets were the following: COVID, COVID 19, #covid, #covid-19, #covid19. After going through the data with the above keywords, we realized there were already a lot of tweets which reflects people's opinion towards the vaccine. We wanted to focus on a more holistic view of the pandemic situation in the eyes of the social media users, so that we can achieve optimal salient topics to analyze with. We also included hashtags because often people would use hashtag when talking about coronavirus related topic instead of using the word itself.



| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | **Category** | **# of positive** | **# negative** | **# neutral** | **# total** | **%** |
| 2 | Vaccine | 59 | 83 | 19 | 161 | 16.26 |
| 3 | Restrictions | 34 | 73 | 18 | 125 | 12.63 |
| 4 | Symptoms | 29 | 56 | 19 | 104 | 10.51 |
| 5 | Impact | 44 | 144 | 35 | 223 | 22.53 |
| 6 | Demography | 19 | 97 | 45 | 161 | 16.26 |
| 7 | Other | 16 | 30 | 51 | 97 | 9.8 |
| 8 | Testing | 13 | 36 | 16 | 65 | 6.57 |
| 9 | Variant | 12 | 20 | 22 | 54 | 5.45 |
| 10 | **Total** | 226 | 539 | 225 | 990 | 100 |

Figure 1: table partitioning all the tweets in their categories and by their sentiments

To make sure that each tweet is unique, we removed retweets, quotes, and replies from our gathering processes via Twitter's standard parameters for search queries.

In addition, the primary language of query was defined as English using the "lan" parameter of the search endpoint.

Finally and most importantly, Twitter's API also has built in annotations based on tweet context for categories like News and especially COVID-19. This was passed into the query using the "context" parameter with the News domain of 123 and the unique annotation id of the COVID-19 pandemic. This greatly reduced the likelihood of returning irrelevant tweets.

## Methods

The TF-IDF algorithm was used to analyze the most recurring words of the eight different topics. The top 10 words in each category with the highest TF-IDF scores will be selected for further analysis. A python script was written

that returns the number of positive, negative and neutral tweets given a topic and a word. This way we can later analyze which word is most likely to generate more positive or negative response towards a specific topic (See figure 6 for reference).

From a data collection perspective, key decisions such as which keywords to use were instrumental in the shaping of the dataset. Unexpectedly, an online campaign was occurring at the same time as data collection that lead to irrelevant tweets from multiple users or bots. Therefore, the term "tortilla blanket" had to be excluded via the negation operator ("-") in the API Query to account for this unexpected interference.

We ensured that the topics that are chosen are not subjective terms. And When classifying data into different topics, there are some tweets that might have overlap between two topics. A good example of that was tweets that talk about the symptoms of the coronavirus, while it can certainly put it under the category for symptom, it can also be placed under the impact category of how the disease has affected the person's life. In such circumstances, the priority is usually given to symptoms as it is more specific than impact. The ambiguity for the team was about whether impact should be included as a topic as it certainly seems to be very broad, but considering the nature of the tweets seen during open coding as some tweets fall right into this category: tweets that is not specific to any other category and is related to a change in a person's life is considered under this category.

An other category was also included because some tweets just really don't fall in any of the categories above, and having impact as a category limits the number of tweets that are part of the others category.

## Results

After conducting an open coding process for around 250 tweets, we concluded with the following topics, and each topic chosen has clearly defined reasoning and rules that can be applied during annotation. :

- Restrictions:
  For restrictions, it is about COVID-19 related regulations and rules that are applied to individuals or groups that limit their behaviours, such as forcing people to wear masks in the public area, limiting the number of customers per store, and etc.
- Demography:
  Demography consists of anything COVID-19 that affects the society as a whole, that sparkles discussion about the government or anything related to race/ethnicity for instance topics that involve racism.
- Impact:
  Impact consists of any tweet that mentions about how the pandemic has had an effect on their personal life or their mental states.
- Vaccine:
  Vaccine is about any tweet that talks about the vaccine and its side effects. Tweets that talk about symptoms from the vaccine we classified under the vaccine category instead of symptoms.
- Variant:
  Variant is anything related to the COVID-19 variants, particularly the recent Omicron variant and concerns surrounding that. This category was chosen for tweets discussing testing for the Omicron variant, over the testing category.
- Symptoms:
  Symptoms means any tweets related to the symptoms of COVID-19 and its variant.
- Testing:
  Testing means any tweet that relates to COVID-19 testing.
- Other:
  Other would classify any tweet that does not fall in the previous 8 categories. We made sure during open coding that the topics selected were objective and discernible from the data.

Result of the top 10 most important words from eight topics, sorted from the highest TFIDF score to lowest are as following:

- Restrictions:
  "rules", "flight", "lockdowns", "flying", "sit", "bla", "sitting", "jim", "sing", "risky"
- Demography:
  "gmt", "coronavirusupdate", "coronaviruspandemic", "coronavirus", "government", "died", "total", "deaths", "australia", "chaos"
- Impact:
  "team", "union", "died", "philaunion", "players", "proud", "psychological", "issue", "ill", "profits"
- Vaccine:
  "shot", "booster", "boosters", "shots", "vaccinated", "unvaccinated", "reactions", "mrna", "unvaxxed", "jabs"
- Variant:
  "omicron", "african", "variant", "omnicron", "coronavirus", "cost", "strain", "rise", "stocks", "milder"
- Symptoms:
  "throat", "breathe", "symptoms", "smell", "fever", "taste", "lungs", "goodness", "headache", "dry"
- Testing:
  "tomorrow", "tested", "antigen", "test", "results", "kinda", "ur", "antibody", "absolute", "hey"
- Other:
  "assess", "amish", "corvid", "deer", "george", "killed", "dems", "centers", "seh", "ig"

Observing Figure 2, around 22 percent of all the tweets talk about the impact of COVID-19 on their daily life or mental

health, which is the most engaged topic. Followed by vaccine which has a 16 percent engagement rate. The least engagement topic was variant, which stands around 5 percent.

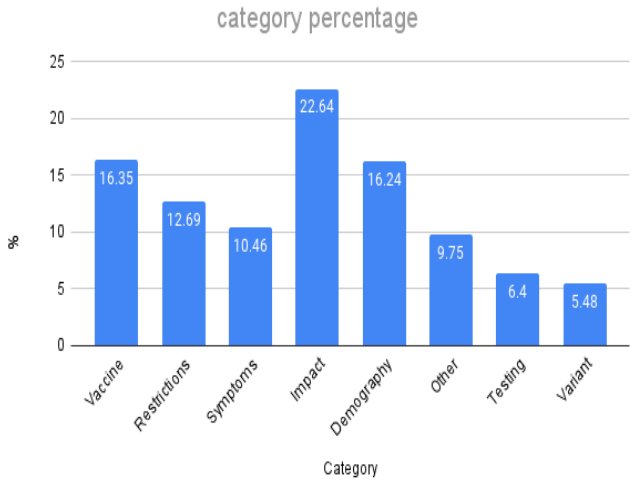


Figure 2: Percentage of tweets in each topic over the total number of tweets

Figure 3 partitions the previous graph by the number of positive, negative, and neutral response to each topic. Tweets in the impact category peaks in terms of negative sentiment at 144 tweets. Despite having the highest frequency in terms of number of tweets, impact category does not have the most number of positive tweets (19.5 percent). We can see that the vaccine category contains 59 tweets which outmatches the 44 positive tweets from the impact category. In fact, 36 percent of all vaccine related tweets seem to be positive or optimistic in their response to the vaccine, and around 51 percent tweets are negative towards it. According to figure 3 and 4, the topic of vaccine has the highest percentage of positive sentiment amongst all positive sentiments at 26.1 percent, and only the third highest in terms of all negative sentiments at 15.5 percent falling behind topics related to impact and vaccine.

A surprising remark is that more tweets seem to be neutral about the variant of COVID-19 than the number of tweets that have a negative sentiment towards it, and by far the less engagement topic of the eight at only 5.48 percent in the category percentage, in spite of the fact of the trendiness of the omicron variant.

## Discussion

From our results, We see that although the general response towards the vaccine is mostly negative, it certainly does not bring as much negativity as the social political issues due to COVID-19 nor the pandemic's impact on people's life. And most importantly, it has the highest supporting number among the topics, which means people might not be as upset in taking the vaccine as for example how the pandemic has affected their life. We can also see from Figure 6, that the

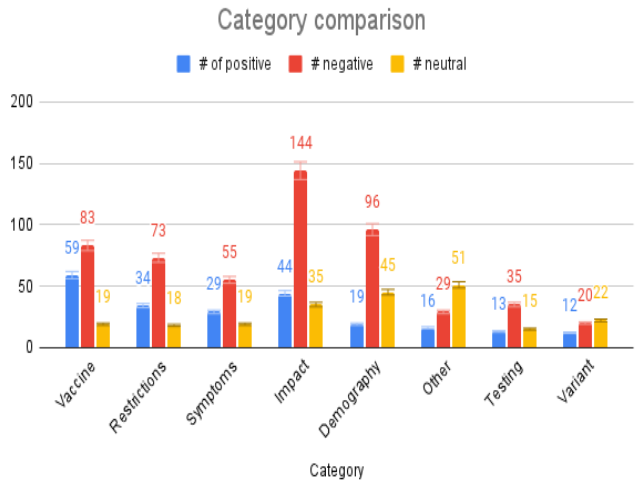


Figure 3: Topic engagement,Category Comparison, and sentiment analysis

word "vaccinated", "booster" in the list of characterization for the topic of vaccine generate almost as many number of positive tweets than negative ones, which signifies that people have their opinion on both end of the spectrum regarding the vaccine, at least based on the data collection we gathered.

Despite the omicron variant, not a lot of people seem to be very concerned about the topic of COVID-19 variant with respect the other topics in spite of the Omicron variant. While the word is indeed the most frequent word in that topic, it comes last in terms of topic engagement, which is quite contradictory with what we expected, given that there are a lot of news report about the Omicron variant causing another lockdown and further restrictions.

Most tweets were very engaged on how COVID-19 has had an impact on their personal life, and it is predominantly negative as shown in the result section, which is quite predictable considering that the pandemic situation has affected people's daily life in a very negative way and twitter is largely used for people to voice out their complains and distress. It can also be reflected through the results of the most frequent used words in that category, we can see that the words "killed" and "died" are among the most frequent uncommon words. In fact of tweets were related to a family member passing away due to the symptoms of COVID-19 and such tweets contribute quite significantly to the negative response of the impact category. Another three very frequent used words were "union", "philaunion" and "players", which might be explained by the incident on December fourth where 11 players from the Philadelphia Union were not able to participate to the Major League Soccer Eastern Conference Final because of COVID-19 protocols.

Restrictions is heavily related with the words "rules", "flight" and "lockdowns" as there are a lot of new cases aris-

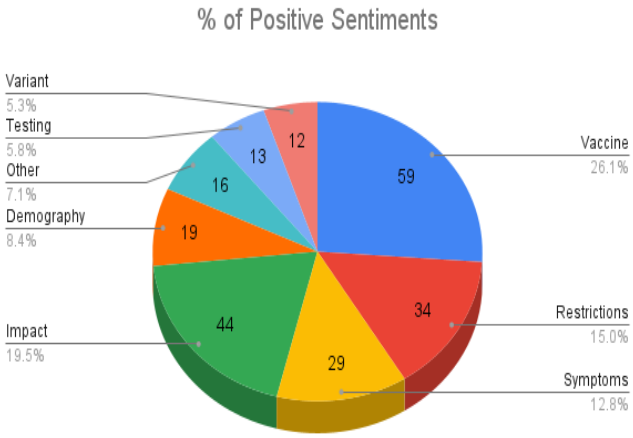


Figure 4: partition of the positive sentiments of the tweets

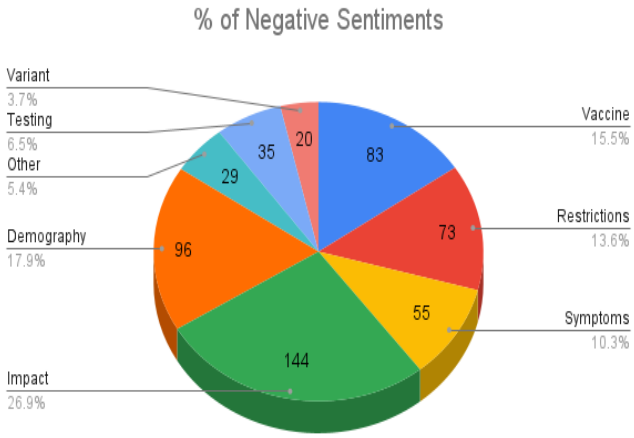


Figure 5: partition of the negative sentiment of the tweets

ing which means that regulations and lockdown restrictions are implicated as well. Hence explaining the frequency of those words.

When it comes to the symptoms of COVID-19, it is easy to see that most people still experience the most common symptoms of the coronavirus disease. As it is reflected through the TF-IDF of the symptoms' topic, people have been using mostly "taste", "breathe", "throat", and "smell" in their tweets, which are problems that corresponds to to the most common symptoms of the disease, which explains where most of the negative sentiment in this topic comes from. Some positive tweets in this category do contain those words as well but mostly they were tweets that expressed the user's relieve after not having COVID-19 despite of similar symptoms.

In terms of demography, most of the negative tweets were related to the government or a specific political representative, in how they have failed addressing proper measures to cease the contagion of the disease. To illustrate that point, in figure 6, we can see that the word "government" has appeared in 12 negative tweets and 0 positive tweets within the demography category. There were also multiple discriminatory tweets towards a certain ethnicity/race that were associated with spreading the disease. The neutral tweets are mostly just reports of new cases on a national level, as we can see the word "coronavirusupdate" is associated with 9 neutral tweets within Demography related tweets, and 0 in the rest of the sentiments.

Testing has more neutral sentiment than positive because some people are about to be tested and don't know what to expect. But a lot of them have a very negative response rate for going to be tested. The negative sentiments also come from people who are upset of testing positive.

As for others, tweets that don't make any sense or does not

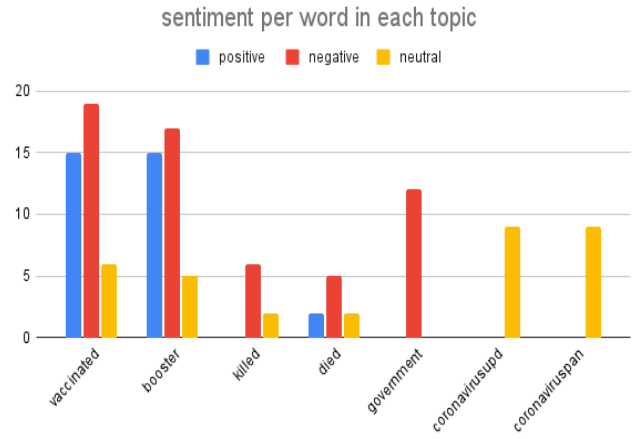


Figure 6: sentiment of tweets by words per topic

fall in any of the above categories are included, others only occupied 9.73 percent of our entire data set, which means that most tweets were able to be classify into a category that is analytical. There were a lot of posts that have typo and some have very peculiar grammar which makes it hard to distinguish the meaning behind the post, even if we try the best to filter our data, such kind of tweets cannot be avoided.

In general, the coronavirus disease has a very negative response rate. According to figure 1, amongst the total 990 tweets, there are 539 negative sentiment tweets and only 226 positive ones, which yields a 54 percent negative response rate, and only an abysmal 23 percent of positive response rate.

## Group Member Contributions

- Yifang: Defining topics of analysis Approx. 350 annotations, Wrote introduction, data section, results section, and discussion section of the report. Generated 2 charts.
- Muzhi: Developing TF-IDF analysis and python script. Approx. 250 annotations. Wrote methods of the report
- Arman: Tweet collection, approx. 400 tweet annotations. Wrote project report data section, editing of report. Generated rest of charts.