Comp598 Final Project
on Covid in General

Written by

Yining Wang, Mcgill ID 260844811, yining.wang3@mail.mcgill.ca
Yuxuan Tian, Mcgill ID 260836395, yuxuan.tian@mail.mcgill.ca
Zhiming Zhang, Mcgill ID 260840709, zhiming.zhang@mail.mcgill.ca

Introduction

In recent years, the prevalence of Covid has brought tremendous impacts on the society. Not only there are massive influences on politics and economics from a macro perspective, covid has affected our daily activites to a great extent as well. We conducted a study to understand what are the discussions happening around Covid on twitter, especially concerning what are people's opinions on taking vaccines.

To be more specific, for this project we aim to find salient topics discussed around COVID and study what each topic primarily concerns based on a collection of 1000 tweets scaped from Twitter within a 3-day window. After manually reading through some of the collected tweets we formed 6 topic categories and characterizing each topic by computing 10 words in each topic with highest tf-idf scores. We also discussed what are the distribution and engagements for each topic, and how positive/negative responses to the pandemic/vaccination has been.

Summarizing our findings, negative responses are playing the majority role in all the topics, statistics are computed at the later part of this report to reveal the sentiments towards specific topics.

Data

In this section we would describe the data set and all the relevant statistics for this projects.

The data collection and filtering are conducted in the following process:

- Creating a Twitter developer account and an app, and by building Twitter's API, Tweepy, we set the fundamental tools we need for collecting data.
- Initially, we collected 1600 tweets within a 3-day window. At this phase, we have encountered several difficulties, mainly in the process of collecting the data itself, and adjusting the format and quality of the tweet contents.
- We did some processing on the raw data and removed all the duplicated contents, which we would discuss in more detail in the 'Methods' part of the report.

API Setup

After setting up the authentication for the Tweepy API, we started to collect tweets for the first time. However, we were getting an error message saying 'Rate limit exceeded.'. To fix this, we added an additional argument for the API set up so that we automatically wait for rate limits to replenish.

```
api = tweepy.API(auth, wait_on_rate_limit=True)
```

Above is our code for the API set up.

Data Collecting Phase

- Query Setup

  Our query keywords for collecting tweets were straight to the point. As we all know, Covid has always been a hot topic since the epidemic. In order to collect data with high association with our topic, and not to include unrelated content, we designed our query keywords to be as simple and straightforward as possible.

  ```
  query = "(covid) OR (canada covid) OR (Covid-19)"
  ```

  Above is our query keyword.

- Specifying Language

  However, with the query keyword as indicated above only as a parameter, it was not enough if we wanted to get English content only. We added the parameter 'lang = en' in order to filter out all the non-English contents.

- Collecting Full Content of Tweets

  Settling down the query keywords and language parameters, we ran a few test trials to collect small amounts of tweets to see if everything was going as we expected. Unfortunately, we found that most of the contents we were getting were truncated. We only got to see the first 140 words per tweet. We thought this was fatal for later annotation because we would not be able to understand the tweet with truncated contents. We added the parameter 'tweet_mode = extended' when calling the searching function, but this was not helping us too much.

  ```
  cursor =
  tweepy.Cursor(api.search_tweets,
                q="(covid) OR (canada covid) OR (Covid-19)",
                lang="en",
                tweet_mode='extended').items(1600)
  ```

Above is our function call used to collect tweets.

```
"RT @chrischirp: THREAD on UK covid situation,
Omicron and what's
happening...

TLDR: Omicron increasing very fast and we are
getting increa..."
```

Above is the example of a typical truncated tweet.

Through our efforts, we discovered that for a retweet, the text would be truncated to 140 characters even after including the parameter 'tweet_mode = extended'. To have access to the full-text content, we need to look at the retweeted_status field of the JSON response.

- Specifying Total Number of Tweets to be Collected

  We were first tempted to collect exactly 1000 tweets as required but through subsequent processing (which would be discussed in the follow-up part of the report), we found that after removing duplicate tweets and meaningless contents, what was left was not enough for us to analyze. Therefore, we have increased the amount of raw data that needs to be collected initially. After several testing trials, we decided to collect 1600 tweets as raw data, and after some later process, we would have 1063 tweets left for analysis.

  At the end of this step, we obtained a dataframe that has all the texts of tweets as columns. We did follow-up processes and analysis on this dataframe.

- Data filtering

  We tokenized each tweet content and made string comparisons after removing stopwords in order to drop all the duplicated tweet contents. Some detailed explanation is provided in the 'Methods' part of the report.

Methods

- We have performed some processing on the raw data to remove duplicate parts and meaningless content. We have 1063 tweets left at this stage.
- We have read through 200 tweet contents and summarized 6 topics after discussion, experiments and deliberation.
- Annotation was conducted by our group members on all 1063 tweet contents after topics were specified. Positive, neutral, negative sentiments were also labelled.
- Tf-idf scores were conducted for all the words that belonged to each topic. 10 words in each topic with the highest scores were selected for future analysis.

Data Processing and Filtering

The raw data has a lot of duplicates as well as some meaningless information, thus filtering and processing are necessary.

We wrote a script to do so. With the help of the Python Pandas library, we converted the texts into lower cases and then tokenized the texts of tweets by splitting the string by spaces. Moreover, with the help of the Python NLTK library, we conducted lemmatization on the data that grouped together the different inflected forms of a word so that they can be analyzed as a single item. However, in later part we discovered that this would cause some weird behaviors on words-grouping. For example, the word 'nonsense' becomes 'sense' after processing. We decided to still conduct lemmatization because without this, there would be a decent number of repeated words with only a slight differ on tense showing when computing top 10 words in each topic with highest tf-idf scores.

Notice that our contents contained some user-specific information which did not help for our purpose of understanding covid on Twitter. Therefore, we replaced all the usernames with the string 'user' and also all the URL hyperlinks with the string 'url' as well. Moreover, all the stopwords were excluded from our text contents. Symbols, retweeted signs('RT'), emojis, and non-English words contained in sentences were also excluded at this stage.

We created a new column namely 'tweets' in our dataframe, containing the list of words that we filtered out as strings. We then run the drop duplicates function provided by the Pandas library on that processed column. We were getting a cleaned dataframe with 1063 tweets in this way.

Topic Selection and Annotation

After filtering, we took a subset with 200 tweet contents and tried to generalize topics by reading through all of them. Two members of our group came up with 6 topics in alignment and we started to annotate the data. One of us actually conducted a double annotation on the dataframe.

However, while labelling we were opposed to more information and we realized that our topics were too detailed, which led to a result that some topics only contained a handful of content. Therefore we generalized and merged some topics that shared some overlap to some extent into one larger topic.

More details in the topic selection part are discussed in the 'Results' part of this report.

Computing Tf-idf Scores

In the data filtering phase, we have already produced a cleaned text column that contains only words that would take into account. Therefore, we only need to compute the word counts for each word and compute the tf-idf score based on their formulas provided in class. We also wrote an additional line of code to sort the scores and generate the top 10 words within each selected topic.

Results

Topic Definitions

After collectively reading through each of the 200 tweets, we came up with six possible topics as follows:

Topic Definitions with Examples		
Topics	Definitions	Examples
Policy	Discussions on the government's regulation correspond to the covid pandemic, also political related contents	Colorado Gov. Jared Polis is a Democrat and he just admitted COVID "is over." He says Colorado will encourage vaccinations instead of mandating masks: "The emergency is over. You know, public health [officials] don't get to tell people what to wear; that's just not their job."
Vaccine	People's opinions on various vaccines of covid	@DailyMailUK Had one pfizer had a heart issue afterwards. Had Astra Zeneca as my second. I will not have another Pfizer or moderna I'd rather take my chance with covid. Why didn't we approve a whole virus vaccine rather one that just looks at the spike proteien
Health	Health related contents including covid test results, symptoms, and mental health under pandemic.	Today is 7 weeks since my kid got covid. He is not better and it feels like every doctors appt we go to ends with. It's shocking to me how little urgency there is to help a kid who is in chronic pain, went from running and playing to barely able to get out of bed".
Life	Discussion on changes of daily life, such as online school, WFH, personal finance, social gathering activities and travel restrictions.	The business of being a business coach has changed dramatically. How can you grow your consulting practice in the midst of Covid-19? Listen here https://t.co/zK8rprDsFw for three action items you can use today. #business #consultant
Covid itself	Studies and researches on covid, including newly found variants and etc.	COVID-19: Omicron is by far the most worrying form of coronavirus we've seen so far - but we shouldn't be too afraid #SmartNews https://t.co/8fzMMilccu
Others	Non related contents.	"Covid proved this beyond doubt: "It is dangerous to be right in matters on which the established authorities are wrong." —Voltaire"

Table 1: Topic Definitions with Examples

- Policy: government regulations
- Vaccine: any discussion on vaccination
- Health: symptoms, recovery, mental health
- School: school regulations(online schedules, exams arrangements, etc) corresponds to the covid situation
- Travel: travelling restrictions due to covid
- Covid itself: variants, researches and studies on covid

However, as we went on to process the whole dataset, we discovered that there were only a really small number of tweets that were categorized under the topics 'Health' and 'School'. Moreover, a lot of tweets were about how people's life changes after covid, for example, the change in personal income, comments on having parties during covid and etc. These tweets were actually nowhere to label according to our current schemes of topic selections.

Being bottlenecked by the coverage of our topic definitions, we decided to create a new topic named "life" that captures citizens' general living conditions under COVID-19. The category was also extended to include events such as personal finance, work, and social gathering. We then modified and regrouped our classification hierarchy, mapping original labels "travelling" and "school" under the category of "life". We finalized our topic selecting into two main categories, one being anything policy or politic-related, and the other representing the public viewpoint of the virus in general, including the impact of COVID on their lives. Among the general opinions, we separated the most popular topics (vaccine, health, COVID itself) and performed individual studies of those sub-topics. The rest of the reviews related to daily life are categorized into one, and we performed TF-IDF to further investigate what were commonly discussed.

Table 1 below shows our selected topics together with some typical examples.

We then started to annotate our dataset. With everyone annotating with great caution, we also labeled each tweet tied with one of the sentiments: positive, neutral, or negative for future topic engagement analysis. What is worth talking about is we all noticed that there was a noticeable amount of negative sentiments labelled in comparison with positive sentiments. Therefore, we had an expectation for our results to come with a huge disparity in content between positive and negative sentiments at this phase.

Topic Engagement

- Topic Distribution

  Below on Figure 1 we presented the topic distribution chart. We calculated the proportion of each topics amongst the 1000 tweets by using the formula:

$$number\ of\ tweets\ per\ topic \div total\ number\ of\ tweets\ collected$$

Overall statistics on Topics and Sentiments							
Categories	Policy	Vaccine	Health	Life	Covid Itself	Other	Total
Positive	20	46	34	30	28	11	169
Negative	78	71	67	111	40	46	413
Neutral	95	67	64	85	47	58	417
Total	193	184	165	226	115	115	1000

Table 2: Overall topic distribution and sentiment engagement table

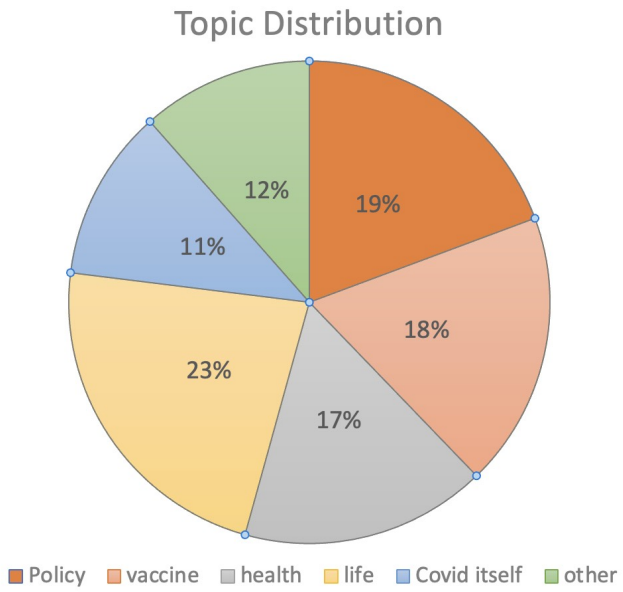


Figure 1: Distribution of Selected Topics

As the table indicated, there were nearly one-fourth of the tweets were about people's daily life activities during Covid. What came next was the policy topic, which was taking a proportion of 19%. Notice that the proportion of tweets that were discussion based on vaccination was very close to the policy topic, with 18% of the contribution.

Not to mention that in order to distinguish and categorize each piece of tweet contents, we labelled tweets to the topic 'other' with great cautious. It was still quite surprising that the proportion of 'other' contents was the second smallest, with the contribution of 12%. Those contents about studies and researches on covid itself had the least contribution of only 11%.

- Topic Sentiment Engagement

  Generally speaking, we were getting similar numbers of tweets that were labelled neutral and negative, while tweets with positive sentiments was the minority part.

  With the aim for achieving the sentiment responses for each topic, we calculated the distribution using the formula:

$$number\ of\ tweets\ labelled\ with\ a\ certain\ sentiment \div number\ of\ tweets\ per\ topic$$

The result is presented in Table 3 below.

Distributions of Sentiments Among Topics			
Categories	Positive	Neutral	Negative
Policy	10.36%	49.22%	40.41%
Vaccine	25.00%	36.41%	38.59%
Health	20.61%	38.79%	40.61%
Life	13.27%	37.61%	49.12%
Covid Itself	24.35%	40.87%	34.78%
Other	9.57%	50.43%	40.00%

Table 3: Distributions of Sentiments(Positive, Negative, Neutral) Among All Topics

In general, the percentages of neutral and negative response are approximately similarly distributed among all the topics, which contributed around 40% respectively. The positive sentiments only counts around 10-20% for each topic.

To be more precise, the topic has most negative responses is those tweets about people's daily lifes, with nearly half of the tweets under this topic are labelled negative. In contrast, the most positive responses came from the topics about vaccination and the studies and progresses on covid itself.

- Topic Characterization We characterize each topic by computing the top 10 words in each category with the highest tf-idf scores, using the formula:

$$tfidf(word, topic, tweet) = tf(word, topic) \cdot idf(word, tweet)$$

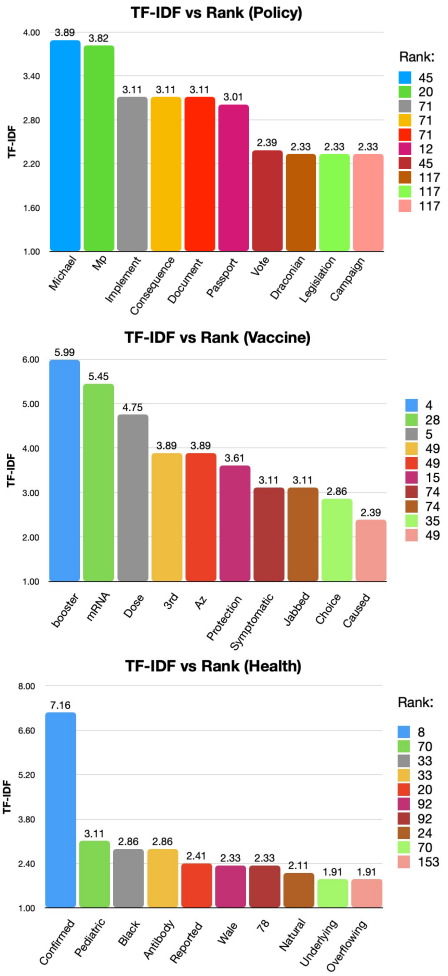
$$tf(word, topic) = frequency\ of\ the\ word\ in\ the\ given\ topic$$

$$idf(word, tweet) = \log \frac{\#_{topics}}{\#_{topics\_contain\_tweet}}$$

Table 4 below lists the top 10 words scoring in tf-idf in decreasing order. We can see those words listed were actually making sense to describe each topics. Further interpretation would be presented in the Discussion part of the report.

Top 10 tfidf words	
Topics	Words
Policy	michael, mp, implement, consequence, document, passport, vote, draconian, legislation, campaign
Vaccine	booster, mrna, dose, 3rd, az, protection, symptomatic, jabbed, choice, caused
Health	confirmed, pediatric, black, antibody, reported, wale, 78, neutral, underlying, overflowing
Life	nonsense, shop, price, customer, produce, poverty, enemy, space, proof, accept
Covid Itself	delta, confirmed, involve, omicron, carry, disability, becoming, transimissible, infect, adding
Other	mug, poor, policy, sentenced, fraud, honest, gate, complained, manson, looted

Table 4: TF-IDF scores



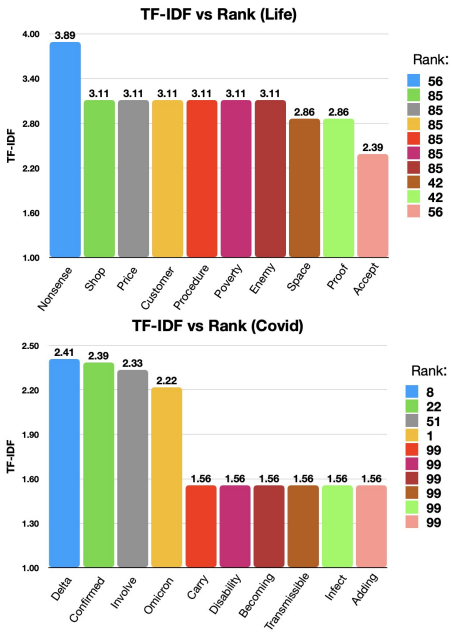


Figure 2: The above figures shows the top 10 words with highest *TF-IDf* score, alongside with their word-count *Rank* in the legend

Discussion

In this section, we will interpret the relationships among categories based on the computed result in detail and demonstrate some typical examples of our dataset as to why they behave in certain ways.

The prevalence of COVID over the years had a massive impact on people's lives – it consists of the largest proportion of negative tweets among all categories, and only 13.27% shows positivity, as indicated by the "life" category of the sentiment analysis. One potential reason for this phenomenon is that COVID has exhausted people's economy. The most discussed keywords such as "nonsense", "price", and "poverty" (ranked 1st and 2ed respectively) all reviewed the financial struggles that people have been undergoing. A few tweets shared their experiences of losing their jobs while their family members were in the ICU struggling for their lives and needed money for immediate treatments.

The government policies and regulations might also have contributed to the financial crisis. Michael Portillo, the ex-Tory MP of England, claimed regulations such as stricter measures such as mask-wearing mandates as "Wholly inconsistent, no one could possibly believe they would be effective" [1], and the cost of maintaining lockdown and isolation weighted over their benefits. Citizens also did not seem to have faith in the current policies. The positive tweets for COVID-related government issues and policies have the lowest satisfaction rate of 10.36%. They supported Michael's standpoint, so claimed that "Tories never lie about anything", making Micheal's name placed 1st in the TF-IDF analysis. The objections against the mandatory regulations

might also form a casual relation with a large number of confirmed cases as indicated in the “health” TF-IDF result, notice that there are still some critics who raised questions about the existence of COVID, claiming that it was merely a made up story. With that in mind, those people who do not believe in COVID would not bear any toleration on public regulations against the spread of the virus, which may also contribute to the result of negative responses.

Forbidding major safety measures will most likely accelerate the spread of viruses, especially during the outbreak of omicron, a variation with increased transmissibility.

As a result, vaccination has the largest ratio of positive tweets. Booster shots, 3rd dose, mRNA, and Az(Astrazeneca vaccine) are widely discussed attributed to the severeness omicron has brought. As time passed, people who were initially skeptical about the effectiveness of vaccines are now finally convinced as more and more real-life examples show the benefits of vaccination.

The new variant brought more confirmed cases, this explains why the word ‘confirmed’ was rated high in the TF-IDF lists for the topic health. More confirmed cases, more regulations promulgate to protect the *public*, more restrictions on daily activities, thus more negative responses overall. We can see that these topics are all related to some extent.

We now discuss some of the limitations of our methodologies. TF-IDF score calculates the amount of information a word contribute to a particular topic. From the results shown in Figure 2, the outcomes with top TF-IDF scores resembles the frequently-used words under each topic closely, as one can see (from the ”Vaccine” category, for example) that words such as ”booster”, ”mRNA”, and ”Dose” are closely related to vaccination. As the result reviews a positive correlation between the frequency of a given word and its corresponding TF-IDF score, the figure also lucidly demonstrated had also made a clear distinction between the two concepts – words that obtained the highest TF-IDF scores are usually not ranked highest in word count, indicating that some words are overused to give any meaningful insight into the topic itself. Though TF-IDF provides a relatively fair measure of information by detecting the occurrence of a given word over the the entire corpus, it also have severe drawbacks – Take the distribution of the word ”vaccine” as an example (Figure 7), although the ratio of its occurrence in the *Vaccine* category has demonstrated its sufficiency in representing the group, it is invalidated simply because the word ”vaccine” is so commonly used that it was covered by all categories (its only occurrence in the ”*other*” group managed to zero out the entire TF-IDF score.)

Also, most of the data our labelling can be intrinsically biased due to the precise/narrow topic definition. For example, the the tweet ”LoL COVID is a scam” can be labeled as both [”other”, negative] or [”covid”, positive] completely depending on how we interpret it. Since we have defined the ”covid” category to be related to the virus itself, such as COVID variants/research breakthrough in relevant fields. One can either categorize it as a negative ”not COVID-related” tweet since it does not reveal any facts(virus structure, fatal rate, etc.) about the virus itself.

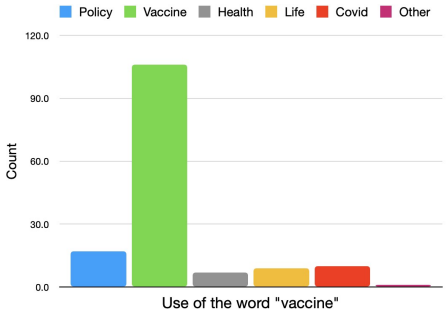


Figure 3: Count of the word ”vaccine” in each category
.

Group Member Contribution

Yining Wang:

- Data collection and annotation, computed Tf-idf scores for each word.
- Wrote introduction, data, methods section of the reports.
- Conducted the result tables.

Yuxuan Tian:

- Data annotation.
- Wrote Result, Discussion section of the report.
- Conducted the TF-IDF graphs.

Zhiming Zhang:

- Data collection
- Data annotation
- Topic distribution graph
- Topic definition chart

References

[1] Schiavone, A. 2021. Michael Portillo erupts at ’preposterous’ plan B as he slams Covid rules inconsistency.