COMP 598 Final Project
COVID in Canada

Written by Beyza Yildirim, Liyun Huang, and Yolanda Ma[1]

[1]McGill University Department of Computer Science

Introduction

After collecting tweets on Twitter within a three-day window, we were able to gather some important information regarding the discussions that were happening around the COVID-19 pandemic and vaccination among the English-Speaking communities. We concluded that there were 6 topics that were the most salient among the tweets we collected: "Virus", "Symptom", "Policy", "Cases", "Impact", and "Vaccine". "Vaccine" is the most discussed topic among all, covering 44.4% of the tweets we collected. Within this topic, "Pfizer" and "Moderna" are the most frequently occurring words, which reflects the fact that these two brands are widely used in North America. Many people like to share their side-effect experiences after vaccination and their opinions on the different vaccine brands. Besides talking about vaccination, the public also likes to share their opinions on politicians' policies on the pandemic and COVID-related mandates on Twitter. With various restrictions imposed by the government and political discussions sparked by the pandemic, "Policy" was, without a doubt, the second most discussed topic in our finding. We also found that, overall, the public's responses towards the pandemic and vaccination were mostly negative. Only 17.6% of the tweets in our dataset were positive, while 32.5% of the tweets were labeled as negative. The fact that the topic "Impact" had 3 positive tweets out of 58 tweets reflected that the public had an overall negative experience during the pandemic. People expressed disappointment in the economy and many discussed their deteriorating mental health due to the pandemic. One noteworthy finding was that, among the 444 "Vaccine" related tweets, there were 26.6% tweets reflecting positive sentiment and 30.6% tweets reflecting negative sentiment. This showed that the public has a moderate hesitancy toward COVID vaccines. While many people expressed support toward vaccination, a large group of people still felt skeptical towards the efficiency and long-term effects of COVID vaccine.

Data

In order to understand the discussions the public has around COVID-19 on social media and learn about people's opinions on the pandemic and COVID vaccination, our team collected tweets from Twitter within a 3-day window.

Our team encountered an issue when we applied for the Twitter API. Each team member sent an application, however, we waited for a week but had not heard anything back about the API approval. Therefore, we used an alternative library called snscrape[1] that allowed us to collect data without the restrictions of Twitter API or Tweepy.

To ensure that the tweets are in English, we set the language field to English when we collected the tweets. Since only tweets that are related to the pandemic or vaccination contexts should be considered in our study, we gathered tweets that mentioned COVID and vaccination-related words. Every tweet in our dataset contained one or more of the following keywords: 'COVID', 'coronavirus', 'COVID-19', 'vaccination', 'vaccine', 'Pfizer', 'Moderna', 'Johnson Johnson'. The first three keywords are common words that people use to discuss the pandemic or the COVID-19 virus; the fourth and fifth keywords are usually used when people talk about the COVID vaccine, and the last three keywords are the brand names of the three most widely used COVID vaccine in North America. By collecting tweets related to one of these keywords, we can assign a COVID or vaccination-related topic to each tweet and analyze the engagement of these topics later on.

Each tweet contains a tweet ID – an identifier that uniquely distinguishes itself from other tweets, text – content of the tweet, and URL – the link to the original tweet, which allows us to revisit the post on Twitter in case the text is difficult to read when we annotate our data.

To factor in the likelihood of collecting retweets, tweets with the same content, or tweets that contain one of the keywords but are unrelated to the COVID or vaccination contexts, we collected 1,350 tweets instead of 1,000 tweets so that we could filter out these tweets. It was important to remove these tweets from our dataset because they would impact the analysis of our results later on. Repeated tweets would lead to over representation of annotations and unrelated tweets do not fit into topics we developed. We collected 450 tweets every 24 hours within the three-day window from December 2nd, 2021 to December 4th, 2021. We filtered out retweets and tweets with the same content by checking if the tweets contain the same text. After filtering, we

had 1,236 tweets remaining in our dataset. Then during our data annotation process, to filter out tweets that were unrelated to the contexts, such as advertisement tweets that used one of our keywords in their hashtags but promoted for unrelated products, we assigned a topic of "Others" to these tweets. After filtering out these unrelated tweets, we had 1,124 tweets remaining. To ensure that our dataset would only contain 1,000 tweets, we performed a random selection and collected 1,000 tweets from the 1,124 tweets.

Methods

For open coding, we randomly selected 200 tweets from the 1,236 tweets we had after removing retweets and tweets with the same content. We went through the 200 tweets and initially developed following 7 topics and provided definitions for them:

1. Variant: Different COVID variants, such as Omicron and Delta
2. Recovery: Recovery and long-term effects from COVID
3. Cases: New cases death statistical reports
4. Impact: Social, cultural and economical effects caused by COVID, the impact of the pandemic on society
5. Vaccine: Benefits/effects of the COVID vaccine and general discussions around vaccinations
6. Restrictions: COVID-related restrictions imposed by the government and other COVID-related mandates
7. Anti-vaccine: Opinions against COVID vaccine After that, we divided posts into 3 parts and each one of us annotated one part.

During our data annotation process, we realized there were some problems with our initial list of topics. Our topics were specific and detailed, however, many tweets in the remaining dataset did not fall into any of these categories. To make sure that every tweet would belong to a topic, we changed several topics while broadening the definitions for some topics.

There were some tweets that discussed coronavirus, but they were not related to some specific virus variants. Therefore, we changed the topic Variant to Virus to include general discussions around coronavirus. We also found some tweets that discussed the symptoms while having COVID, so we changed our second topic from Recovery to Symptoms.

Some users tweeted about cases of COVID where their family members or friends contracted coronavirus, however, the definition of our third topic Cases would not allow these tweets to be categorized. To make the topic more generalized, we decided that Cases should also include discussions of individuals contracting the virus.

We also changed our sixth topic Restrictions to Policy because we realized there were some tweets related to political decisions on COVID-related situations, such as different political parties' opinions on certain COVID restrictions and a country's decision on choosing a particular brand of COVID vaccine. Therefore, we chose to use the word Policy to include political discussions around COVID-related issues.

At last, we removed the seventh topic because we realized people's opinions on COVID-vaccine were already covered in the sentiment analysis. After a series of modifications, we finalized the topics and their definitions:

1. Virus: Discussions around coronavirus
2. Symptom: Symptom while having COVID and long-term effects after recovering from COVID
3. Cases: Statistical and first-hand reports on COVID cases
4. Impact: Social, cultural and economical effects caused by COVID, the impact of the pandemic on society
5. Vaccine: Benefits/effects of the COVID vaccine, vaccination status, and general discussions around vaccinations, such as the booster vaccine
6. Policy: COVID-related policy/restrictions imposed by the government and COVID-related mandates, such as social distancing measures and mask mandates.

After annotation, we calculated the TF-IDF scores for each topic and selected the words with the highest value with the following formula:

TF - IDF score:

$$TF(term|topic) \cdot log(\frac{numberoftopics}{numberoftopicsthattermisusedin})$$

Results

In order to develop our topics, we have conducted an open coding of a subset of 200 tweets from our finalized version of the dataset. At the end of this process, we were able to finalize our topology on the 6 main subsets shown in Table 1.

Table 1: Topic definitions

Topic	Definition
Virus	Discussions around SARS-CoV-2, its variants Positive Example: "Dying COVID-19 Patient Recovers After Court Orders Hospital to Administer Ivermectin #alert #COVIDIOT #epochnews #love #PrayTogether" Negative Example: "You'd think we'd have learned to take #COVID19 variants very seriously, but we again did too little, too late and now it's out in the community in #Australia: Omicron in children under 5: Kids disproportionately hospitalised by variant #auspol #COVID19Aus"

Symptom	Symptoms seen due to COVID-19 and the long-term effects of the virus after recovery Positive Example: ”Bombshell alert. So I actually think there is a silver lining here and this may signal the end of Covid-19, with it attenuating itself to such an extent that it’s highly contagious, but doesn’t cause severe disease. That’s what happened with Spanish flu” Negative Example: ”I got covid and the worst part abt is 1. No taste or smell 2. NO GYM BRO LIKE WTF”
Cases	Statistical and first-hand reports on COVID cases Positive Example: ”Skip Bayless, Stephen A. Smith. and all LeBron Haters. LeBron James tested Negative for COVID-19 and cleared to play Basketball on Friday against the LA Clippers” Negative Example: ”@IBSteveBMe1 @AngelicoDeb @AdamKinzinger That’s mean nothing compared to all this peoples that died because his lies about the covid-19... United States reached its latest heartbreaking pandemic milestone Friday, eclipsing 700,000 deaths from COVID-19”
Impact	Social, cultural and economical effects caused by COVID-19, the impact of the pandemic on society Positive Example: ”@BashirMohamed I’m sorry, its not right. Covid should’ve taught the importance of global efforts in health care, and reminded us of the common vulnerability in all humankind. Instead those with the most are fighting among themselves imagining threats that don’t exist and ignoring ones that do.” Negative Example: ”I’ll tell you what. This Covid-19 situation has changed a lot of people. I see a lot of family issues and mental illnesses growing. The social experiment is working.”
Vaccine	Opinions on the effects of the COVID vaccines, vaccination status of twitter users, and general discussions around vaccinations, such as the booster vaccine Positive Example: ”Got my free Moderna booster shot today and I couldn’t be happier for this little glimpse of universal healthcare.” Negative Example: ”@SpaghettiOrgy @wopizza @Martyupnorth_ So the Pfizer data says 1223 people are known to have died from the vaccine from Dec 1 until Feb 28, 2021. I guess that’s another conspiracy theory proven true. Excited to inject it into your children yet?”
Policy	COVID-related policy/restrictions imposed by the governments, and COVID-related mandates, such as, but not limited to; social distancing measures, vaccination and mask mandates. Positive Example: ”COMMENT: Govt move to tighten regulations against Covid-19 is smart and well-timed” Negative Example: ”@danmccay While you work to find the right balance, 602 Utahns have died from Covid related deaths in October and November. Blood on your hands for prohibiting mandatory testing”

After labelling the tweets with their topics, we have conducted a TF-IDF analysis in order to determine the most relevant words in the collection of tweets per topic. The top 10 words with the highest TF-IDF scores per each topic are shown in Table 2.

Table 2: Top 10 words with highest TF-IDF scores

Topic	Top 10 Words with Highest TF-IDF Scores
Virus	”omicron”, ”variants”, ”strain”, ”travel”, ”lab”, ”omicronvarient”, ”believe”, ”causes”, ”difference”, ”viruses”
Symptom	”severe”, ”pretty”, ”lungs”, ”thanks”, ”suggest”, ”term”, ”flu”, ”common”, ”players”, ”shot”
Cases	”lebron”, ”active”, ”fatality”, ”james”, ”confirmed”, ”detected”, ”india”, ”negative”, ”tested”, ”deaths”

Topic	Top 10 Words with Highest TF-IDF Scores
Impact	”economy”, ”stock”, ”lot”, ”past”, ”mandates”, ”else”, ”seen”, ”supply”, ”stocks”, ”straight”
Vaccine	”booster”, ”shot”, ”dose”, ”pfizer”, ”vaccines”, ”arm”, ”hours”, ”moderna”, ”johnson”, ”status”
Policy	”mandate”, ”proof”, ”employees”, ”house”, ”mandates”, ”congressional”, ”vaccination”, ”religious”, ”fda”, ”countries”

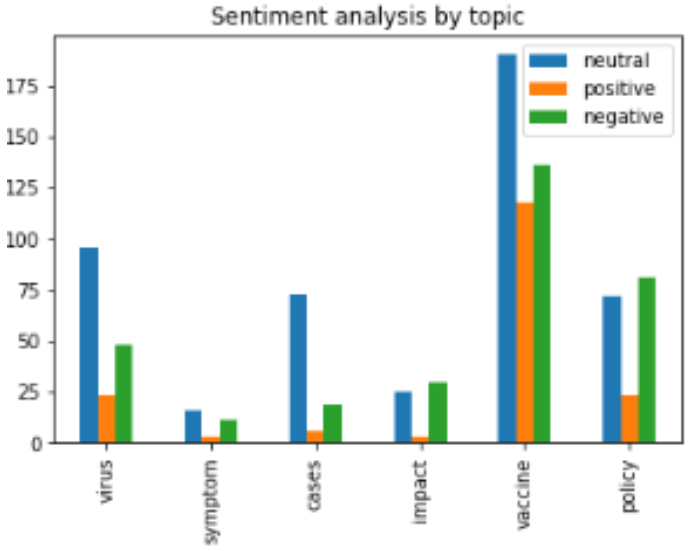

Figure 1: Sentiment analysis with respect to topic

The sentiment analysis results are visualized in the bar plots given in Figures 1 and 2. We observe that overall, with having 472 tweets and almost half of the portion of the entire dataset, the neutral tweets lead the sentiment

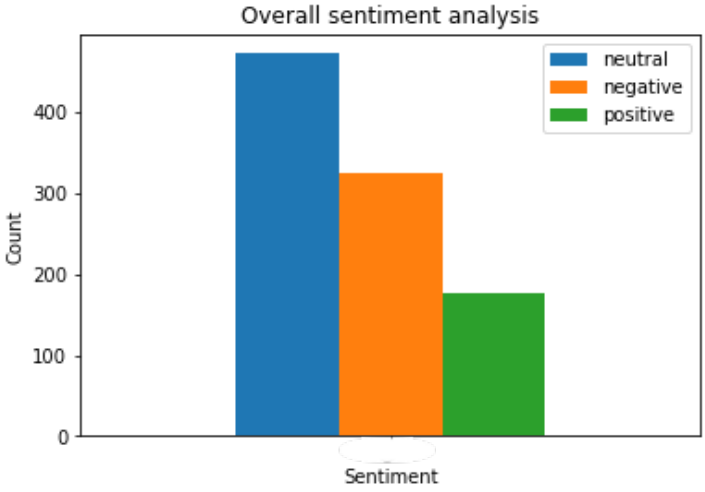

Figure 2: Overall sentiment analysis

analysis. We also see that the number of negative tweets are almost twice the number of positive tweets, indicating the dominance of negative sentiment over positive in the discussions made in Twitter on the topics of COVID-19.

As we go into the sentiment analyses per topic, we notice that the overall trend of neural tweets leading the discussion changes for ”Impact” and ”Policy”, where number of negative tweets dominate over both neutral and positive tweets. Especially the lack of positive tweets for ”Impact” compared to the negative tweets catches our attention. We also observe that although positive tweets have the lowest numbers across all topics, the sentiment seems to catch up the number of negative tweets on the topic of ”Vaccine”.

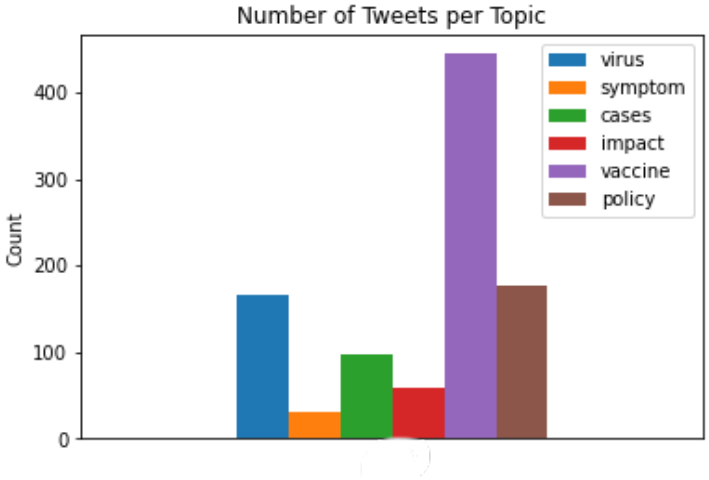

Figure 3: Number of tweets per topic

The results for the number of tweets tweeted per topic are visualized in the bar plot given in Figure 3. We observe that most of the tweets were on the topic of ”Vaccine”. With

444 tweets, this is nearly half of the dataset we have. This number is followed by ”Policy” and ”Virus”, with having 176 and 167 tweets respectively, which are pretty close to each other in terms of popularity. We observe that the number of tweets tweeted on the topics of ”Cases”, ”Impact” and ”Symptom” are not very high, with being 98, 58 and 30 respectively.

Discussion

To analyze the data, intuitively, we can first compare the occurrence of each topic to find the popularity of topics. “Vaccine” has a mention of 444 times, which is more than twice as much as the second most occurred topic, “Policy”. This is expected since the current major event related to COVID is the vaccine. The topic “Symptom” has the least number of mentions of only 30 tweets in this category. This only accounts for 3% of the total tweets we collected.

Looking into the most popular category ”Vaccine”, judging from the word counts, ”pfizer” and ”moderna” appear to be the two most frequent words with word count numbers of 169 and 151 respectively. It makes sense since they are the two most widely used vaccine brands. However, they both occur in the categories ”Virus”, ”Cases”, ”Impact” and ”policy”. Therefore, they ranked lower in the set of 10 words of highest tf-idf scores in the topic ”Vaccine”. the word ”booster” is the most frequent word measured by tf-idf scores. It is well-known that some cities in North America has already started to offer the booster dose to public or to a certain group. We do expect to see different reactions to this new measure. As a result, this result integrates well with our prior knowledge on COVID.

Taking a closer look into the 10 words in each category with the highest tf-idf scores, we can observe that some words appear to be frequent in more than one category. For example, the word ”mandate” is frequent in both topic ”Policy” and ”Impact”. It makes sense since that the word ”mandate” must be widely used in policy and restrictions regarding COVID, and these policies impact on social, cultural and economic aspects of our lives.

One interesting finding we observed is that the word “lebron” has the highest tf-idf score of 16.1 in the “cases” category, and the word “james” also happens to appear frequently in this category with the fourth highest tf-idf score of 7.7. We looked into the relevant tweets and investigated the reason behind it, and it turns out that NBA star LeBron James was tested positive for COVID. He went through additional tests later on with negative results, and that allowed him to return to the game[2]. His tests gave conflicting results according to the news, which we believe is what triggered some more discussions on Twitter. It makes sense since that the breaking celebrity news is always a trending topic on social media.

Additionally, this article was released on December 2nd, and the words ”lebron” and ”james” occurred only on the tweets we collected from Decemeber 2nd and December 3rd. From our data, there’s no tweets discussion on this on Decemeber 4th. This indicates that breaking news is time-sensitive content on social media. It triggers hot discussions as nowadays the users always tend to stay updated. However, this type of content usually has a shorter life span. I believe this finding can be a useful piece of information for the marketers.

The sentiment analysis also provides insightful findings. It has the least number of positive tweets in all the topics. The total number of positive tweets is 176, which is 17.6% of the total tweets. There is a total number of 325 negative tweets, which almost doubles the number of positive tweets. The rest are neutral tweets, accounting for around half of our data. In general, it seems like the public shows more negative emotions towards COVID than positive sentiment.

However, it is worth noting that for the category ”Vaccine”, the percentage of positive and negative sentiment is 26.6% and 30.6% respectively, which is a lot closer than the average percentage difference between positive and negative emotions. It somehow implies that the public has higher acceptance of COVID vaccines compared to their attitudes towards other COVID related issues, but the proportion of positive sentiment may still not be sufficient to increase the vaccination coverage to a high enough level.

Another observation is the topic ”Impact” has the least percentage of positive tweets. There are only 3 positive tweets out of 58 tweets falling in this category, accounting for only 5.2% of tweets in this category. It is well suggested that the impact of COVID on society of all aspects are mostly seen as negative by the general public.

Through analysis, we have learned that there might be approaches that can be taken in order to gain a more comprehensive interpretation of the data.

1. Due to the given time constraints, single annotation is used for this project. This can introduce possible bias due to the coder’s personal point of view. Human coding can also bring potential manual error. A double annotation would produce a less error-pruning result.
2. From the Top 10 Words List with Highest TF-IDF scores, we can observe that some words appear more than once in different forms, such as ”mandate” and ”mandates”, and ”vaccines” and ”vaccination”. If we are able to classify them into one word, we should be able to obtain a more constructive words list with highest TF-IDF scores.

The methods above are some adjustments we might be able to implement if the situation permits. In general, We are confident that our current result does provide a reliable insight on the present discussions about COVID on social media.

Group Member contributions

- Liyun Huang worked on data collection, open coding, data annotation, Introduction, Data, and Method sections of the report.
- Yolanda Ma worked on data annotation, data analysis and the Discussion section of the report.
- Beyza Yıldırım worked on data annotation, data analysis, calculation of TF-IDF scores, and the Results section of the report.

All group members contributed to the write-up equally.

References

[1] Martin Beck. "How to Scrape Tweets With snscrape". In: *betterprogramming* (Dec. 3, 2020). URL: https://betterprogramming.pub/how-to-scrape-tweets-with-snscrape-90124ed006af (visited on 12/13/2021).

[2] Associated Press. "Lakers' LeBron James clears COVID-19 protocols, expected to return vs. Clippers". In: *Sportsnet* (Dec. 2, 2021). URL: https://www.sportsnet.ca/nba/article/lakers-lebron-james-clears-covid-19-protocols-expected-return-vs-clippers/%22,%20urldate%20=%20%222021-12-13.