COMP598 Final Project

Group member
Yao Chen 260910483 (yao.chen5@mail.mcgill.ca)
Mengyu Chang 260901444 (mengyu.chang@mail.mcgill.ca)
Boyu Han 260899328 (boyu.han@mail.mcgill.ca)

Introduction

Coronavirus disease 2019, also known as COVID-19, is an infectious disease caused by the SARS-CoV-2 virus. After the first case has been identified in December 2019, it has been spread enormously across the world and continuously impacted people in the everyday life.

Although COVID-19 has limited in-person social interactions, people are connecting online more than ever by using social media. For many, social media not only play a critical role in keeping friends and family connected during times of forced separation, but also has become a lifeline to the outside world. Especially as people look for ways to remain connected and entertained. There are tons of discussions around COVID-19 on social media. In this project, we discovered the main topics discussed around COVID and analyzed primarily concerns about each topic by looking at the top 10 words with the highest tf-idf score. We also compared relative engagement with those topics by their corresponding frequency. Finally, we discovered what sentiment to the pandemic and vaccination has been based on response.

This main process is following. First, we collected and cleaned the data posts regarding COVID-19 from Twitter within a 3-day window from Nov 12 to Nov 14. Then we analyzed the 1000 posts by manual annotation. After first looking through 200 randomly selected posts, we developed six salient topics, which are vaccine, policy, finance, cases updated, lifestyle, and health care. We then annotated the rest by the topics and sentiment. With the annotated data, we computed tf-idf score and got the top ten words under each topics. In addition, we made some visualizations base on data to get clear and concrete insights.

Data

The source data is collected from these three days 2021-11-12 , 2021-11-13 and 2021-11-14. Our original data set contained 4200 tweets, almost 1400 tweets for each day. We used the following keywords to filter the tweets from twitter, they are Covid, Vaccine, Vaccination, Pfizer, Moderna, Janssen, Astrazeneca. And after some operations ,our final collected data set contained aroud 1400 tweets.And we randomly choose 1000 tweets from 1400 tweets as our analysis data set and do further research on it.

We know the analysis data set is very important ,it will directly affect our result. If our data set is not accurate about covid-19, then our result will not be reliable. So below is the details about how we get these 1000 tweets and make sure they are all about covid-19.

At beginning, we decided to collect as much as possible tweets to guarantee we have enough data to use.Therefore, we prepared to get 600 tweets that mentioned Covid,200 tweets that mentioned Vaccine, 200 tweets that mentioned Vaccination, 100 tweets that mentioned Pfizer, 100 tweets that mentioned Moderna, 100 tweets that mentioned Janssen and 100 tweets that mentioned Astrazeneca for a day.As a result, it should be 1500 tweets in total for each day. But when we actually collected the tweets, unfortunately we found for some keywords like Janssen and Astrazeneca, there are only 40 to 60 tweets in a day, so the data set for a day contained only 1400 tweets and the source data contained 4200 tweets.

After we got the source data, we randomly choose 1000 tweets as our analysis data.And again randomly choose 200 tweets from the 1000 tweets to do annotation.But when we do annotation on these 200 tweets, we found there are some retweets, some tweets mentioned almost the same things and some tweets without mentioning the keywords.For the first two thing(retweets and tweets with similar text), we know this maybe occur sometimes, but for some tweets without keywords, we have no idea.So we checked the data again and found this may due to the tweets include not only the text, but also include some other headers which may contain some keywords and our collecting method focus on all headers of the tweets so that the keywords may do not show on the text of the tweets. These inaccurate and duplicated data is quite a lot (almost 50 tweets in 200) and we cannot ignore them.Thus we decide to do data cleaning to make the tweets in our data set unique and accurate.

1. Here, we first did some duplicate removal operation, including detecting retweets and remove them. We try to

compare each tweet and remove the tweets with same content. But this only remove the retweets, there still left some tweets with extremely similar text and we will solve this later.

2. Second we double checked the text of the tweets to make sure at least one keyword is indeed included in the content(Since we also find some tweets we collected by the keywords but strangely no keyword in their content).
3. Third we did artificial judgment to check if the text contained the keywords is related to the covid-19 (Since for some keywords like Vaccine and Vaccination, they may be also included in some tweets that talk about HIV and so on).
4. At last, after above three data cleaning steps, our data set only contains around 1800 tweets.Then we sort the tweets by their texts,so that we could easily find the tweets with similar text by looking around a tweet and remove them.

After these operations, we finally got our final accurate data set which contains almost 1400 tweets. And the 1000 tweets for further analysis are all comes from this.

Methods

We attempted three strategies of collecting our source data.

Our first strategy is to directly access data from Twitter API by using the Request library in python. The original version of the keyword list was designed before we actually requested data. They are ”Covid”, ”Vaccination”, ”Moderna”, ”Pfizer”, ”Astrazeneca”, ”Janssen”, respectively. Since there were limited monthly tweet cap usage for our developer account, we determined to gather 1000 tweets as our raw data. It was designed to check whether our keyword list is efficient and desired. After the very beginning check, we found three main problems for collect our raw data:

1. A large proportion of ”Janssen” tweets were not discussing topics regarding the approved vaccine ”Janssen”, instead the person who called ”Janssen”. We imposed the keyword list such that the selected ”Janssen” tweets should contain both ”Janssen” and ”Vaccine”.
2. The collected tweets have the bias on the ”post time” as they were all posted around midnight on that day. To extract tweets randomly over a specific day, instead of setting the start-end date as a whole day, we narrow the range of tweet post time to one hour, and then collect tweets by hour.
3. The majority of tweets lack discussion about Canada, which is opposite to our aim to focus on Canadian social media. Extracting the user location becomes our focal point. Due to the restriction on extracting the location of the user, we utilize our second strategy.

Our second strategy is to collect tweets by using Python library tweepy. Different from our first method, we collected 1500 tweets as our raw data this time, without bias on post time. Some new improvements were implemented on this stage:

1. Some tweets were not shown with full content. We observed that those tweets were either written in paragraphs or retweets. To address this issue, we set the tweet mode to ”extended” mode, and split all raw data to normal tweets and retweets at the very beginning. If the post starts with ”RT”, it is a retweet, and thus we extract full text from ”retweeted status”. Otherwise, we collect text from ”full text” directly.
2. Moreover, a new keyword ”booster” was added to the keyword list since it appears frequently on our new raw data. However, since our twitter developer account had the restriction on accessing the location of the user. We failed to achieve our goal. And then we carried out our third method.

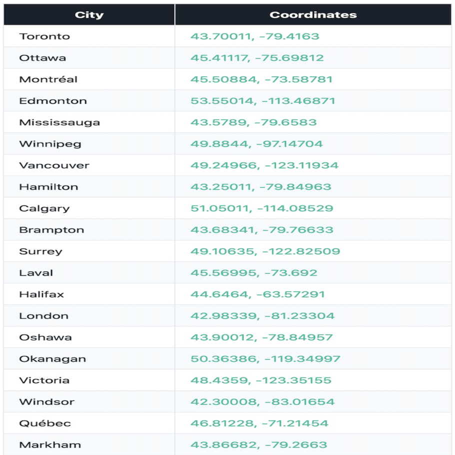
Our third strategy is to collect tweets by using another Python library snscrape which is developed by a GitHub user and this tool is extremely useful to collect the data from twitter. Below is the instruction link to snscrape: https://github.com/JustAnotherArchivist/snscrape

To use snscrape, we first need to update our python to the latest version which is python3.10 and then plug in pip3 install snscrape to the command line of our terminal, then we can use the models in snscrape.And for exploring data on twitter, there is a special model named snscrape.modules.twitter.This model contains a bunch of methods to select and filter tweets from twitter which meets all our needs(including to check if the tweets comes from Canada).So we first create a query which contains all the selected rules such as keyword, date, location and language.Here we focus on how to determine the location(since other selected rule is almost the same as the first two strategy, then here we left them out).

The query of snscrape.modules.twitter allow us to give an range that indicate the post location of a tweet in this format (x, y, d) as an selected rule.Then snscrape.modules.twitter will automatically call the twitter API to check if the post locations of the tweets are in that range (the first two parameter x and y are the geometric coordinate of the center location and the third parameter d is the radius of that range) so that we could get the tweets of Canada by setting appropriate coordinate.

The most straight forward way is to pick the center location of Canada as the geometric coordinate (x,y) and the distance between the center location to the the nearest border to Canada as the third parameter d.But this does not work in the way that we thought , since the range is too big and it includes some tweets from other countries.

Then we use another approach.That is reducing the range and query multiple times, so we want to collect the tweets from the cities of Canada.Here we only focus on the big cities (that is the cities with enough population because there are little tweets about covid-19 in small cities).Then we used google and got the geometric coordinates for the big cities.Below is the screen shot of the cities and their coordinates.



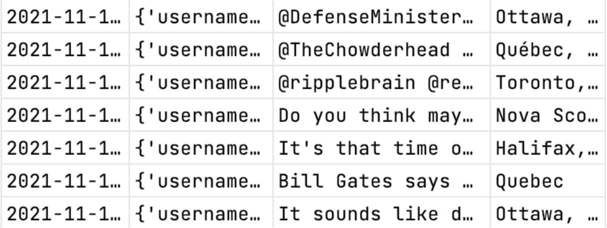
City	Coordinates
Toronto	43.70011, -79.4163
Ottawa	45.41117, -75.69812
Montréal	45.50884, -73.58781
Edmonton	53.55014, -113.46871
Mississauga	43.5789, -79.6583
Winnipeg	49.8844, -97.14704
Vancouver	49.24966, -123.11934
Hamilton	43.25011, -79.84963
Calgary	51.05011, -114.08529
Brampton	43.68341, -79.76633
Surrey	49.10635, -122.82509
Laval	45.56995, -73.692
Halifax	44.6464, -63.57291
London	42.98339, -81.23304
Oshawa	43.90012, -78.84957
Okanagan	50.36386, -119.34997
Victoria	48.4359, -123.35155
Windsor	42.30008, -83.01654
Québec	46.81228, -71.21454
Markham	43.86682, -79.2663

Thus we get exactly the tweets of Canada by this approach.And the remaining thing is to do iterations to get tweets containing different keywords for each day and city. So here, we also import another python library itertools to help us do the iterations.

At last we import python library pandas to create a dataframe to store the details of the tweets.Then we select the certain columns of the original dataframe to form a new dataframe containing only the useful things for annotation and analysis.The certain columns are as follows:

1. date: when the user post the tweet
2. user: including the information of the user such as user id, nickname and so on
3. content: the text of the tweet
4. location: the city where the user post the tweet

   Finally, the data set looks like this:



2021-11-1…	{'username…	@DefenseMinister…	Ottawa, …
2021-11-1…	{'username…	@TheChowderhead …	Québec, …
2021-11-1…	{'username…	@ripplebrain @re…	Toronto,…
2021-11-1…	{'username…	Do you think may…	Nova Sco…
2021-11-1…	{'username…	It's that time o…	Halifax,…
2021-11-1…	{'username…	Bill Gates says …	Quebec
2021-11-1…	{'username…	It sounds like d…	Ottawa, …

Result

After collecting and cleaning posts from Twitter, we make decision on topics together. We first randomly selected 200 posts from the clean data and manual annotation on those. We looked through the first 50 posts together, generalize the main insight of the contents and had five choices for the topics. Then we annotated the rest of posts separately and compared the results together.

While comparing the results, we did have different annotations on ambiguous cases. In this situation, we looked at the posts again and tried to persuade with each other. For example, some posts contain statistics on vaccine. One of us thought it is statistics topic, while others thought it should be vaccine topic. As this condition happened several times, we then decided if the post is vaccine-related, it will all goes into vaccine category, regardless whether some other topics are also covered or not. This process is to make sure we all have a clear understanding of the meaning each topics. During the annotation, we also discovered that ”Healthy Care” are mentioned a lot in posts, so we also added this topic.

Finally, we categorized posts into six topics as following:

1. Vaccine (v): This topic is about vaccine-related posts including authorized vaccines, booster dose update, people’s response after getting a dose, different opinions about vaccination.

2. Policy and politics (p): This topic is about COVID-related policy and politics, including the local and global response towards COVID.

3. Finance and economics (f): This topic is about effects of COVID on Canada economy, pricing, and personal finance status.

4. Cases announcements (c): This topic is about COVID-related statistics report or trends.

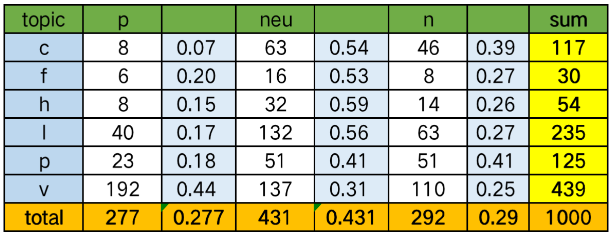
5. Lifestyle (l): This topics about Canadian citizens’ lifestyle due to pandemic and COVID-related community event.

6. Healthy Care (h): This topic is about Healthy Care system, medical support, hospitalization and treatment related news.

When we manual annotated on the rest of posts, multiply topics may be covered in one posts. We then make a decision base on the central idea of the post. On the other hand, one post may also hard to identify into six topics. We then base on the key words or the link mentioned in the post to get the topic.

Besides coding for topics, we coded each post for positive, neutral or negative sentiment by using 1, 0 and -1 respectively. When the posts does not have very clear sentiment, for example, the statistics report, we will always mark it as neutral. In some cases, although the sentiment of post is not neutral, but base on the content, the attitude is not caused by COVID. Then we will also mark it as neutral. This is to make sure all the marking of response is toward the pandemic or vaccination, taking away the influence of other factor.

The following table is presented the topic engagement and frequency of positive or negative sentiment on each topics. The float number in blue represents the ratio of sentiment to total number of posts.



topic	p		neu		n		sum
c	8	0.07	63	0.54	46	0.39	117
f	6	0.20	16	0.53	8	0.27	30
h	8	0.15	32	0.59	14	0.26	54
l	40	0.17	132	0.56	63	0.27	235
p	23	0.18	51	0.41	51	0.41	125
v	192	0.44	137	0.31	110	0.25	439
total	277	0.277	431	0.431	292	0.29	1000

From the table above, we can draw the following conclusion:

- The number of the positive posts is 277, neutral posts is 431 and negative posts is 292
- The Number of different topics from largest to smallest is Vaccine (439), Lifestyle (235), Policy and politics (125), Cases announcements (117), Healthy Care (54), Finance and economics (30). People cares more about the Vaccine and Lifestyle but less people think about the Healthy Care and Finance and economics.
- The topics with the highest proportion of positive post is Vaccine that is 0.44 and the topics with the lowest proportion of positive post is Cases announcements that is 0.07
- The topics with the highest proportion of negative post is Policy and politics that is 0.41 and the topics with the lowest proportion of negative post is Vaccine that is 0.25
- The neutral posts always account for the largest proportion among the six topics except the Vaccine.

In order to characterize each topic, we computing the top 10 words in each category with the highest tf-idf scores using python. In this procedure we used the stop word list and expanded it, since we found some words that does not make sense at the first time we directly computed the tf-idf score.After filtering words by the expanded stop word list, the result is as following and the words are sorted in the decreasing order of tf-idf scores:

- Vaccine (v): moderna, pifzer, shot, astrazeneca, booster, anti, vaccination, medical, dose, kids.
- Policy and politics (p): airborne, wrong, protocol, government, house, austria, onpoli, kenney, wave, mask.
- Finance and economics (f): pay, pfizer, finances, economy, stocks, strategy, covid, health, amp, vaccine.
- Cases announcements (c): active, deaths, london, quebec, middlesex, unit, sunday, rate, scotia, confirmed.
- Lifestyle (l): team, game, play, hit, season, masks, lost, players, time, lineup
- Healthy Care (h): capacity, ivermectin, banning, testing, nursing, hospitals, flu, patients, icu, cold.

Discussion

We will first discuss each topic characterization in detail.

Vaccine(v) has the largest numbers of tweets which is 439/1000 among all the topics with 192 positive posts, 137 neutral posts and 110 negative post.The positive posts is mainly about the benefits for people to prevent the covid and some positive speech to encourage the public to get vaccination. Inversely, the negative tweets is mainly posted by some people who do not want to get vaccination and criticize the vaccines.The neutral post is almost all about what the vaccines do and how they works, that's somthing like the popularization of vaccines and posted by some exports.

Then move onto the top 10 most frequency words(by tf-idf) appeared in this topic.They are moderna, pifzer, shot, astrazeneca, booster, anti, vaccination, medical, dose, kids. We can see moderna, pifzer and astrazeneca are in the first four words, while for Janssen, it even does not appear in the top 10 words, this indicates the first three vaccines are really famous and popular in Canada, lots of people talk about them and give their opinions, but very few people mention Janssen. For the reason of these, we guess Janssen vaccines cames to Canada later than the other three vaccines.Next, we find shot, booster, dose are also mentioned frequently, this implies the public is concerned about how they will get the vaccination and what about the dose of vaccine.We also find the word anti is always mentioned, after we checked the data, we found it mainly came from the tweets posted by the anti-vaccine people, they think people need not to take vaccines.Last but not least, the word kids is also in the top 10 words.For this, people want to know if the vaccines have some bad influence on their children and whether the kids need to take vaccines or not since COVID-19 mainly affects middle-aged and elderly people.

Policy and Politics has the largest proportion of the negative posts(51/125) whereas very low percentage of positive posts(23/125).This is due to the public complain about some policy make their life become boring.For example, A policy is during the COVID-19, only the people who is fully vaccinated can stay in the restaurant and enjoy their dishes, others should take their dishes away and eat them at home. So this policy is something like making the vaccination mandatory for some people and they complain about the policy limit their freedom. While in our group, we agreed that most of these polices are good to protect people from the covid, we should work together to overcome this particular period.

Move on to the top 10 words in this topics.The words wrong implies some people think certain policy does not make sense and should be canceled. And some places' names appear like austria may be due to the 3 days we chosen to collect the tweets and in that days, the government build some international protocols with other country.At last, we find the word mask, this is reasonable since most of the neutral tweets is about the policy to call on people to wear mask in crowded places and public places.

"Finance and economics" is the least popular topics among the all, with only 3 percent. The number of response about negative sentiment roughly the same as positive sentiment. While most of the sentiment are neutral which takes up over half of the posts. This means that most of the posts are about facts of finance and economics. Among all other topics related to COVID-19, people did not pay too much attention towards this topic.

When analyze the top 10 words, to our surprise, it contains vaccine and one of the brand – Pifzer. Pifzer is also the word with second highest tf-idf score. This shows that people are caring about the market of vaccines, especially about Pifzer. One of the technical term in the list is AMP. AMP is abbreviation for asset management plan. This is a way to understand the overall health and needs of these valuable assets and make informed decisions for the future. It reflects the worries and uncertainties of Canadian citizens due to the pandemic.

The topic ”Cases” has the third largest rate among all, with about 11.7 percent. Slightly over half of the posts have the neutral sentiment. This is understandable since most of the statistics are used to present the updated cases, and this should not have express attitude in them. However, the rate of negative sentiment are much outweighed the negative sentiment by more than 32 percent. This reflects the situation with COVID-19 is not optimistic in Canada, at least among the three day period when we collected the data.

Now move on to the top 10 words in this topic, ”active” and ”deaths” are most commonly mentioned words. It represents the active cases and deaths number happened in Canada. It is very common to used in the updated cases reports. The next three words with high tf-idf scores are three city names: ”London”, ”Quebec” and ”Middlesex”. We can infer that the outbreak cases may happened in this area, so that this places are discussed the most about the case updated.

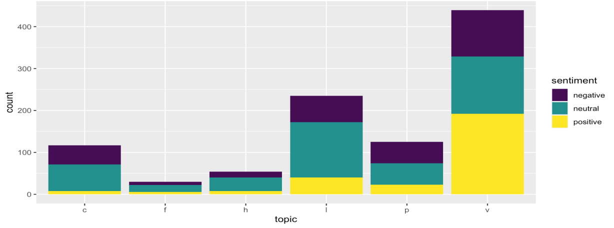
”Lifestyle” is the second most popular topic in our results. Among one thousand tweets, approximately 23.5 percent of them relate to the lifestyle of Canadians during the pandemic. The response with neutral sentiment occupies more than half of all lifestyle-related tweets. Meanwhile, the positive and negative responses accounted for 27 percent and 17 percent, respectively. Along with the continuous battle of the COVID-19 pandemic, people gradually adapt to the new lifestyle during pandemic, since there are large proportions of non-negative responses.

By analyzing the top 10 words that characterize the topic ”lifestyle”, the following insights were found: Most of the lifestyle-related tweets discuss the ongoing and upcoming ball games and their daily life during the pandemic. As we can see, ”mask” have already become an indispensable part of people’s lifestyle. Several public events such as ball ”game” are resuming, ”team” and ”player” has become focused.

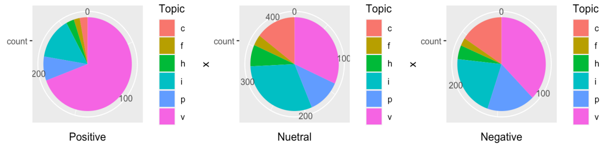
For topic ”Health Care”, there are about 60 percent of health care related tweets involving neutral speech. The negative sentiment shows a higher ratio than positive, with 26 percent.

Among the top ten words in this topic, ”capacity” appears most frequently, with the highest tf-idf score in this topic. It indicated people’s concern about the healthcare system’s capacity to cope with the new wave of pandemic and the coming ”flu” season. People have focused on how the public health agency of Canada is balancing the pressure on the health care system, including the arrangement of medical resources, the treatment for COVID-19 ”patients” and caring for ”patients” with other health issues. Covid ”test” has been mentioned a lot in this topic as well. To combat the pandemic effectively, there are a lot of people on Twitter who recommend those who have influenza or COVID-19 like symptoms to do the test. Moreover, ”Ivermectin” , the second popular word in this topic, is one of the WHO approved medicines which is used to prevent and treat covid. It derives that the prevention and treatment of COVID-19 infection is also turning to a part of people’s concerns.

The bar graph presents the number of posts in each topics, with purple represents the negative sentiment, green represents neutral sentiment and yellow represents positive sentiment.



Three pie chats present the proportion of sentiment for each topics, with positive, neutral and negative sentiment respectively.



Group Member Contributions

Boyu Han:

1. Participate in data collection. Mainly collect the data by snscrapy, filter the tweets by language, date and location, make sure the tweets is about Canada.
2. Do data cleaning, remove retweets, tweets with extremely similar content and tweets without the keywords.
3. Do data analysis, including complete the stop word list to filter the words with no sense and compute the tf-idf score.
4. Participate in data annotation, discuss with other group members about the topics to choose by the 200 sample tweets and manually annotate the rest 200/800 tweets.
5. Participate in writing the report, responsible for the following parts: Data, Method, Result and Discussion.

.

Mengyu Chang:

1. Participate in data collection for one day period and doing the data cleaning for the raw data set.

2. Participate in data annotation. Doing the first 200 sample tweets with group together and manually annotate the 300/800 tweets
3. Participate in writing the report for the parts: Introduction, Result and Discussion.

.

Yao Chen:

1. Participate in data collection and cleaning process. Collecting raw data from Twitter API by using request and tweepy. Design the basic structure of output json. Filter out the duplicated tweets with same content and different links.
2. Participate in data annotation process. Complete the first 200 open coding tweets. Discuss and summarize the topics with details and manually annotate the 300/800 tweets
3. Participate in writing the report for the parts: Method, Result and Discussion.

Reference

A. Bryant, “Ivermectin for prevention and treatment ofcovid-19 infection: A systematic review, meta-analysis, andtrial sequential analysis to inform clinical guidelines,”Journal, 2021.

Twitter, “Getting access to the twitter api docs twitterdeveloperplatform.”https://developer.twitter.com/en/docs/twitter-api/getting-started/getting-access-to-the-twitter-api, Dec 2021. Accessed on 2021-12-10.

CIHI, “Overview: Covid-19’s impact on health care systems.”https://www.cihi.ca/en/covid-19-resources/impact-of-covid-19-on-canadas-health-care-systems/overview-covid-19s-impact, Dec 2021. Accessed on 2021-12-10.