COMP 598: Final Project

COVID in Canada

Aybuke Ekiz, Florence Yang, Keyu Yao

McGill University

Ekiz, Aybuke aybuke.ekiz@mail.mcgill.ca ID: 260973863

Yang, Florence jiashu.yang@mail.mcgill.ca ID:260885843

Yao, Keyu keyu.yao@mail.mcgill.ca ID: 260906814

Introduction

The outbreak of COVID since 2020 has created great chaos and changes in people's lives fundamentally. Living in modern society, posting on social media has become one of the most popular and easy ways for people to express their sentiment and share personal experience and therefore a reliable and quick source for COVID related sentimental and categorization analysis. This project mainly deals with the collection and analysis of the English discussions currently happening around COVID topic on Twitter. Particularly, it is targeted at the salient topics around COVID, what these topics concern the most and their relative engagement, along with the sentimental tendency towards pandemic and vaccination.

The project is designed to collect Covid-related tweets and analyze the most frequent words that people mention under the topics around Covid, and the process involves data scraping, topic developing, feature extraction and topic characterizing. We collected 1,000 tweets within a three-day window from 2021-11-30 to 2021-12-01 that mentioned either COVID, vaccination, or a name-brand COVID vaccine. The collecting of the data is implemented using Tweepy API and the data preprocessing is conducted using Pandas in Python. Conducting open coding on the first 200 hundred tweets, we came up with 8 groups: Vaccine stance, Anxiety, Vaccine side effects, Case/Death updates, Politics, Restrictions, Symptom/Testing and Others. Then, we annotated the rest of the Tweets according to these categories. Among these topics except for Others, we found that "Vaccine Stance "(201 tweets), "Cases/Death updates" (150 tweets) and "Politics" (120 tweets) have proved to be the most popular topics with the highest number of tweets relevant while "Vaccine side effects" being the least discussed topic (39 tweets). When it comes to the sentimental expression, we found out that most of the tweets collected are neutral (512 tweets) but negative tweets (357 tweets) outweigh that of positive ones(131 tweets). Not surprisingly, similar patterns can also be seen within separate topics. For different categories, we also found that "Anxiety/Frustration" is the most "negative" topic with 68.7% of its tweets annotated negative, while "Restriction" is the "least positive" topic, having only 2.5% of its total tweets expressing positive sentiment.

For different topics, we also developed ten feature words with highest TF-IDF scores to better understand what are the main discussions within each category. We found that in general most of the characterizing keywords are highly correlated to the topics but with relatively low TF-IDF scores except for topic "Vaccine side effect" which gives the highest score at 0.466 for word "pfizer".

Data

The data source used for this project is Tweets collected from a three day window of Nov 30th to Dec 1st on a daily basis. Originally we used several keywords during the data scraping process using Tweepy API: "vaccination", "COVID", "biontech", "pfizer", "sinovac", "sputnik" ,"moderna", "johnson" and "janssen". After that we collected around 200 tweets using these keywords to do the topic developing but the result was not as optimal as we expected. Since most of the tweets also involve username,

there was a certain amount of tweets collected which are irrelevant of the COVID but having “Johson” or “Janssen” in the username. However, when looking at the data collected, there were not many tweets that actually talked or mentioned about Johson or Janssen as vaccination brands, so these two keywords did not contribute too much to the relevant data collection, on the contrary, it produced confusion in the data and increased the number of tweets that are irrelevant of the topic of interest. For that consideration, we trimmed these two keywords down and only left the rest 7 keywords during data scraping of the 1000 tweets, filtering all the retweets. At the same time, since there are a lot of official media accounts on Twitter posting news, we noticed that some of them have the exactly identical context only with different https addresses suffixes, that was caused by how we collected the data, since we only collected the last three days of tweets, the same news are highly possible to be reported multiple times, which might make our data less representative. In order to avoid duplication, we eventually collected around 1200 tweets and manually trimmed them down to 1000 tweets after the deletion of identical news tweets.

Except for that we have also collected some of our data from the tag COVID combined with the ones retrieved using covid as keyword. The main reason is that we found a non-negligible amount of tweets collected by the latter method, were not talking about covid or vaccination, but only mentioning the keywords as an addon of the context. For example, we encountered some tweets like “smoking cigarettes is worse than covid” or “the socks I wore before covid”, which mentioned covid but did not really talk about it, only using it as a configurative term or bringing it up as a timeline. Given that our main purpose is to dig around the topics of actual COVID discussion, we ended up using the tweets from tag COVID to dilute the tweets that are “fake” covid-related.

Methods

Data Collection

Data collection was done by twitter API. We set the date of search, which is from 2021-11-30 to 2021-12-1, and our search words. After that, we filter the tweets by language set to English, and filter out the retweets. We saved our data in a csv file for future annotations.

Data Annotation / Categorization

We first decided on doing the characterization on the 200 tweets. While doing open coding, we found it was hard to limit our groups to 8 since there were so many different topics people tweet about. Therefore, we used some of the topics as a general term for including some subcategories. We used the topic “Anxiety/Frustration” as a general title for the Tweets that are about anxiety or emotional expressions towards Covid, which are related to restrictions, health concerns etc. We also included posts about how Covid has impacted the user’s life. We used the category “Vaccine Side Effects” for people tweeting about their personal experiences with the vaccine side effects as well as news articles about vaccination side effects. We used “Vaccine Stance” as a general topic encompassing “anti-vax” tweets, “pro-vax” tweets (tweets that encourage people to get vaccinated), and people’s thoughts about booster vaccines.

What is worth noticing is that, during the period of our data collection window, LeBron was tested positive and there were several discussions related to it. Although seemingly too specific and could reduce the objectivity, we realize that the discussion on twitter over celebrities in general is in fact very common. Within different periods the targeted object could be different but the phenomenon of people talking about celebrities getting covid should not be overlooked. As a result we decided to not filter any tweets related to the discussion of LeBron getting COVID. We also did not create a new category for that either, and instead tagged tweets talking about LeBron as “Case/Death” updates.

We did our annotation by category and sentiments of positive, negative or neutral.

Calculation of TF-IDF Score & Stopwords

For feature extraction and the calculation of TF-IDF, we chose to use the tfidif vectorizer module from sklearn. Originally we used the built-in English stopwords, as a result, we found that all of the topics have “covid” and “vaccination” with rather high TF-IDF scores (around 0.5 to 0.7). The reason behind it is not hard to understand since during the data scraping we put “covid” and “vaccination” in the keywords. However, these words are way too generic for each topic and cannot further deliver information about what each topic is really concerned about. As a result of that we decided to add “covid” and “vaccination” in our stopwords as well. Additionally, we have also added words, such as “https”, “vaccines”, “like”, “just” to the stopwords which appear in the tweets a lot but are less informative compared to other words under our context, and also the numbers like ‘000’ which are used frequently to report covid cases.

Results

Our data were selected into eight topics: Anxiety/Frustration, Cases/death Updates, Politics, Restriction, Symptom/Testing, Vaccine side effect, Vaccine Stance and Others. Figure 1 shows the number of tweets in each category indicating the engagement of each topic. Figure 2 analyzes the sentimental results from each category. Figure 3 shows the top ten words with highest tf-idf scores.

Numerical representation of the number of tweets in each topic:

- Anxiety/Frustration {'negative': 68, 'positive': 8, 'neutral': 23}
- Cases/death updates {'negative': 36, 'positive': 18, 'neutral': 96}
- Politics {'negative': 73, 'positive': 6, 'neutral': 41}
- Restriction {'negative': 39, 'positive': 3, 'neutral': 76}
- Symptom/testing {'negative': 14, 'positive': 15, 'neutral': 45}
- Vaccine side effect {'negative': 16, 'positive': 4, 'neutral': 19}
- Vaccine stance {'negative': 72, 'positive': 50, 'neutral': 79}
- others {'negative': 39, 'positive': 27, 'neutral': 133}
- Total {'negative': 357, 'positive': 131, 'neutral': 512}

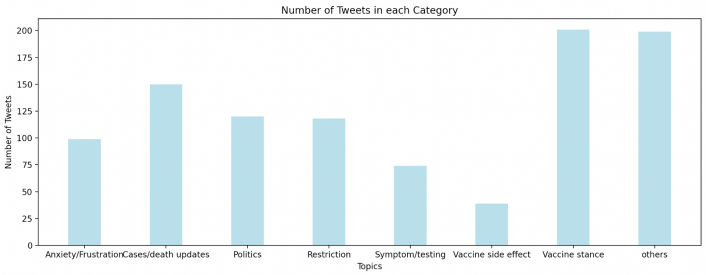

figure 1:topic engagement

Anxiety/Frustration		Cases/death updates	
people	0.296812	omicron	0.427211
know	0.178087	cases	0.427211
long	0.158299	new	0.397405
world	0.158299	variant	0.288119
going	0.138512	york	0.188768
think	0.138512	lebron	0.168897
way	0.118725	died	0.129157
pandemic	0.118725	people	0.119222
work	0.118725	deaths	0.109287
right	0.118725	vaccinated	0.099351

Politics		Restriction	
biden	0.352408	nfl	0.211613
people	0.264306	restrictions	0.195335
trump	0.249622	protocols	0.179057
pfizer	0.117469	unvaccinated	0.162779
joe	0.117469	mask	0.146501
deaths	0.117469	germany	0.146501
million	0.102786	status	0.146501
government	0.102786	people	0.146501
gop	0.102786	rules	0.130223
work	0.088102	fuck	0.130223

Vaccine stance		others	
people	0.336949	people	0.221500
booster	0.327590	vaccinated	0.147667
vaccinated	0.308870	great	0.132900
pfizer	0.224633	nfl	0.132900
unvaccinated	0.131036	time	0.132900
stop	0.121676	know	0.118133
getting	0.121676	health	0.118133
immunity	0.121676	say	0.118133
virus	0.102957	flu	0.103367
fully	0.102957	pfizer	0.103367

Symptom/testing		Vaccine side effect	
test	0.376479	pfizer	0.466498
home	0.155021	moderna	0.219529
long	0.155021	booster	0.192087
testing	0.155021	effects	0.192087
people	0.155021	batch	0.164646
google	0.132875	vax	0.164646
reviewer	0.132875	docs	0.164646
severe	0.132875	deaths	0.164646
effect	0.132875	reports	0.137205
new	0.132875	reaction	0.109764

figure 2: top 10 words with tf-idf scores in each topics

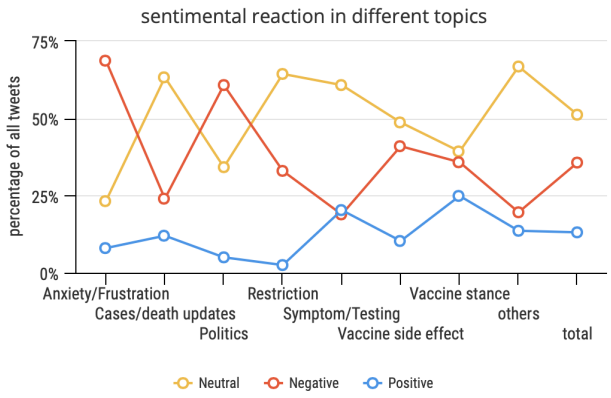

figure 3: sentimental reactions

Average Number of "@" in a Tweet for Each Category	
Anxiety/Frustration	0.03
Cases/Death Updates	0.47
Politics	1.31
Restriction	1.01
Symptom/Testing	0.92
Vaccine Side Effect	1.05
Vaccine Stance	1.57
others	1.16

figure 4: reply tag usage per category

Discussion

According to figure 1 in our result, it is noticeable that the topic “vaccine stance” has the highest engagement. As we reflect this back to our data, a lot of people are tweeting about their personal opinions on vaccination, including a lot of “anti-vax” and “pro-vax” sentiments. Others category is the second highest amongst these eight, which is caused by the limited topics we chose to do the annotations, therefore a great amount of tweets that don’t fit any topic went into “others” easily.

Analyzing the “Vaccine stance” category a bit more, we also see that this category is also the category with the most reply tags (“@”), as can be seen from figure 4. Since the “Vaccine Stance” is the most subjective category among our topics, and very prone to disagreement, we can deduce that the abundance of the “@” tag shows the people arguing back and forth about their opinions on the vaccines.

According to figure 2, the one thing we noticed from the tf-idf table was how much the time frame would affect the topics discussed in the Tweets. For example, in the time period we collected our Tweets (2021-11-30 to 2021-12-01), LeBron catching COVID, Antonio Brown faking his vaccine status to NFL, the Omicron variant and Germany imposing new restrictions for unvaccinated people were all very recents events and they have been reflected in the Tweets. That is a natural and expected phenomenon to occur since Twitter, like all social media platforms, is a fast-paced environment. Being influenced by the recent updates, tf-idf values of “omicron”, “lebron”, “york” (regarding recent omicron cases being detected in new york), “nfl” and “germany” were high. If we were to collect the Tweets in a different time frame, we would be likely to get different words with high tf-idf values regarding the recent updates of that particular time frame.

According to figure 3, the relative negative/positive sentiments of different topics is an interesting discussion topic too. We can for example see that the “Anxiety” topic has the highest negative sentiment, followed closely by “Politics". For anxiety, that makes perfect sense because the way we defined the anxiety topic was “health anxiety, anxiety about restrictions, and COVID’s effect on daily life”. Since these topics are negative by definition, the sentiments of the tweets that would go under these topics were expected to be negative as well. For politics, however, the result was more interesting, because politics, by definition, is not negative or positive, and in our topic definition, we did not have a natural sentiment for the politics. However, most of the tweets under the politics category turned out to have a negative sentiment. This sentiment is also reflected by the fact that “deaths” in figure 2, and also the words regarding the current and also former president of the United States and the Republic Party which are among the highest tf-idf words for the Politics category. All other words for the politics category except “deaths” and “pfizer” are inherently referring to politics, combining with figure 3, Politics has a high

negative sentiment ratio, thus we can infer people blaming the deaths to politics, and complaining about the politics for what they do towards covid.

The category that turned out to be most neutral was the restrictions. It was because most of the tweets labeled as restrictions were the ones updating about new restrictions that will take place. They were either referring to news articles or were tweets from newspaper accounts. However, the negative tweets still outnumbered the positive ones. There were 39 negative tweets, and only 3 positive tweets. The negative sentiment tweets were mostly complaining about the new restrictions being imposed and wearing masks. They also included swear words, which explains the high tf-idf values of the words “mask”, and “fuck”.

We can also discuss the common words among the different categories. We see that the word “long” appears as a word with a high tf-idf value in both Anxiety, and Symptoms/Testing. While this can point to a possible overlap among the topics, it also showcases the people’s anxiety about covid symptoms, and especially about the long COVID.

It is also important to think about our search words. We searched for “Covid”, “vaccination”, and vaccine brand names. While calculating the tf-idf values, we included “Covid”, “vaccination”, and “vaccine” as stopwords, since they appeared a lot in each of the categories and did not offer us any new insight about the topic engagement. We, however, decided not to include vaccine brand names as stopwords, and instead to calculate their tf-idf values to figure out their relative engagement. While we searched for “moderna”, “pfizer”, “biontech”, “sinovac”, “sputnik”, and “astrazeneca”, the only two that appeared as high tf-idf words were “pfizer” and “moderna”, with pfizer’s tf-idf value being double of the moderna’s tf-idf value. We can conclude then, pfizer and moderna being the most popular vaccine brands among Twitter, with pfizer with a tf-idf score around 0.47 while moderna’s is around 0.22, we can simply conclude that the pfizer is about twice populer of moderna at the time frame the tweets were collected.

While some words, like “booster” are self-explanatory, some words like “stop” require deeper analysis and explanation. When we manually go over the Tweets that include the word “stop”, we see that the word is mainly used in two different contexts. One, as an imperative, as in “Stop downplaying #COVID19. Stop saying kids don't get COVID. Stop saying they don't get severely ill.” or “please save our children stop the Covid Vaccine Mandates.” or “You are mistaken. The vaccine does NOT prevent you from getting Covid. It ONLY makes the effect of the virus milder. Also, blaming the unvaccinated for the spread is misleading. The vaccinated can spread the virus also. Please, educate yourself, or stop lying.” In this imperative context, the word stop is used in order to stop doing something to another group of people. Since these tweets belong to the “Vaccine Stance” category, which is highly subjective by definition, and the category in which people share their own personal opinions, this use of stop offers us insight about how people choose to share their opinions about the topic, which is by being aggressive, and often rude. The other major use of the word stop is in a different context, as a synonym for “prevent”. Some examples for this usage are “There is no question the Covid vaccine is imperfect. If it was perfect, it would stop transmission. Therefore, vaccines cause mutations and new variants. So leave the unvaccinated out of it”, “Vaccines won’t stop the spread. It’s a known fact that vaccinated people will still get and spread Covid. In fact, the first US citizen identified as having Omicron was a vaccinated man.”, “I'm in two minds. Being vaccinated doesn't stop you catching or passing Covid to others. I know two former-shielders who recently caught Covid from their fully-jabbed partners. Thankfully, they're both fine, probably because they were double jabbed. If anything, it should be a recent test.” With some examples, it is clear that the first usage of the word stop and the second usage of the word stop are different, and they could have been two different words in a different language. This analysis offers us additional insight on the word contexts.

Group Member Contribution

As a group, our approach to the project was to work on it together as much as possible. At first, we met together to discuss which topics we wanted to choose, and how we wanted to approach the project timeline. After deciding on the COVID topic, we jotted down some tentative dates to meet. In our next couple of meetings, we mostly did the Tweet collection part together. We were getting an authentication and Twitter API time-out error, Keyu worked on those errors and solved them. Just to have some idea about the topics, we tried open coding on the hundred Tweets we managed to collect before getting the time-out error and noted down the possible topics we came up with. Then, once we had our 1200 Tweets collected, Aybuke and Florence met to open code 200 Tweets and enhance the topic descriptions. Once we had our topics set, we started annotating all the Tweets. We split the data evenly into three, everyone did the annotation on their own time, and after the annotation was done, we met again to calculate the tf idf values together. During the same meeting, we discussed what we want to include in the final report and partitioned the report among ourselves. Florence worked on the Introduction and Data. Aybuke and Keyu met together to work on the Methods, Results and Discussions. We also all read over the report together and made the final edits.