COVID in Canada

Li, Yixuan and **Hong, Darien** and **Qian, Grace**
yixuan.li@mail.mcgill.ca
darien.hong@mail.mcgill.ca
grace.qian2@mail.mcgill.ca

Abstract

COVID has had a huge impact on the world and it's citizens; In this report, we present our analysis on the discussions currently happening around COVID on social media in the English language.

In this report, we analyze the salient topics discussed around COVID and what each topic primarily concerns. We also discuss the relative engagement with each topic and how the general sentiment response to the pandemic and vaccination has been. Using a data set that represents COVID discussion on social media in the English language, we conducted an open coding to develop the main topics and annotated with these topics and their overall sentiment. We then characterized the topics by computing the top 10 words in each category with the highest tf-idf scores. We present our results and findings using calculations and graphs.

Data

Our analysis drew on Twitter posts (tweets) as our main form of data. Using a Twitter Developer Account, we made use of the Twitter API to collect 1,000 tweets within a 3-day window, from December 1st to December 3rd. There were 300 tweets collected on the first day, 300 tweets collected on the second day, and the remaining 400 tweets were collected on the 3rd day. All of the tweets collected were in English and to ensure that they related to our COVID topic, we chose the following keywords which we believed would help us collect relevant data. All of the keywords were case-insensitive.

Covid	Covid-19	Coronavirus
Vaccination	Vaccine	Variant
Pandemic	Pfizer	Moderna
Quarantine	Johnson&Johnson	AstraZeneca

Using these keywords, we were able to collect tweets that revolved around COVID and the salient topics discussed around COVID. A setback from using this set of words was that 'JohnsonJohnson' sometimes gave some tweets that did not have anything to do with COVID, and simply contained the word 'Johnson' in it. For example, a tweet about 'Dwayne Johnson' appeared in our data set, but the tweet had no mention of the pandemic. This occurred during our data collection phase since we couldn't ensure that every single tweet we collected was relevant. Instead of replacing each of these tweets, during our analysis, we marked these tweets with the label *irrelevant* so that we knew not to include them in our tf-idf computations. This is a design choice that we decided to keep for our tweet collection process because it is a relevant Vaccine brand; it just meant we had to be more careful during our analysis to catch the tweets that were not relevant for our problem. We also removed any duplicates from identical tweets and retweets to ensure all of the tweets in our dataset were unique.

It is important to note that since our dataset was only comprised of around 1,000 tweets, this is not representative of all of the discussion currently happening around COVID in social media. Using this small subset of data, we were able to conduct a thorough analysis of salient topics, engagement, and sentiment towards the pandemic, which may be accurate in our small sample, but may lose its accuracy when we compare the results to an analysis conducted on a larger and more representative data set. Since our data was collected over a period of only 3 days, the tweets could also be biased and contain information that was relevant during that time period, which may also affect it's representation of the public's true discussion around COVID.

Methods

After collecting the 1,000 tweets for our data set, we conducted an open coding on a subset of the tweets. From this open coding, we were able to develop 6 topics based off what we believed were the main subjects of most of the tweets, including an 'other' topic. The topics we chose to represent the data were the following:

- **public policies** – this relates to restrictions, travel restrictions, government regulations, and mandates (masks, vaccine passports, social distancing, quarantining).
- **vaccinations** – this relates to any discussion surrounding the COVID vaccine excluding specific public policies regarding vaccinations (these tweets would fall under 'public policies'). This topic includes getting vaccinated, side effects, scientific research on vaccines, and solutions to COVID. The reason public policies regarding vaccinations is excluded from this category is to be

able to analyze the sentiments of vaccinations outside of required mandates or regulations. Doing so, we can better analyze Twitter users' feelings about the COVID vaccine's safeness and effectiveness in a way that's isolated from public policies.

- **variants** – this relates to any discussion around current/past variants of COVID, their effects on society, and ongoing research.
- **impact on humans** – this relates to any discussion surrounding the impact of the pandemic/COVID on a human level, including positive/negative test results and cases, deaths, as well as the changes in daily routines and lives.
- **political opinions** – this relates to any political opinions towards the pandemic and the regulations, including sentiment towards the government, people with certain political alignments, and specific politicians. This includes tweets that criticize members political parties and/or their behaviors as well as attacks politicians for supporting certain restrictions, mandates, or policies. These tweets are distinct from 'public policy' tweets because the main focus is on the Twitter user's opinion toward politicians and/or people with certain political alignments.
- **other** – this topic label is for any tweets that do not fall under any of the other topics, but are still relevant to the COVID discussion. For example, some Twitter users tweet about music they have listened to or funny pictures they have found during the pandemic, but they do not talk about the pandemic itself.

As mentioned in the **Data** section, we also had a label *irrelevant* for those tweets that were collected but did not contribute to the discussion around COVID. This was not a topic but instead a label for us to ensure these tweets were not involved in our analysis and calculations. During the open coding, we developed multiple other topics that seemed relevant at the time, such as **covid testing, restrictions, mortality, people with COVID, celebrities with COVID, quarantine, government,** and **politics**. Upon further analysis of the rest of the dataset, it was clear that we could group some of these subtopics into more general topics that would better represent the data. We also decided to add the *other* topic only after we starting manually annotating the data because we realized that there were some tweets that didn't fall under any of the other categories yet still were relevant to COVID in some way. After finalizing the main topics, we proceeded with manually annotating the rest of the dataset, which involved reading through each tweet and choosing which of the topics best represented what the tweet was centered around.

During the manual annotation, we also coded the tweets for sentiment - **positive, negative** or **neutral**. We used some keywords to help distinguish positive and negative sentiment tweets, such as **excited, thrilled, cool, nice, happy, great, love** for positive tweets and **sucks, horrible, terrible, hate, disgusting** for negative tweets. On top of just using these keywords to narrow tweets down, we also read each tweet through to fully understand the sentiment since these kinds of words can be used in different types of contexts and cannot determine a tweets sentiment on its' own. Most tweets that did not express a particular opinion or sentiment were labelled 'neutral'. Majority of the tweets fell under the 'neutral' category as they did not encompass any particular sentiment and may depend on external context to fully understand.

Using the annotated data, we were able to compute a few statistics to get a better idea of the salient topics discussed around COVID and what each topic primarily concerns by calculating term frequency inverse document frequency (tf-idf) scores on the words that appeared within a topic as well as tweets' sentiments. Stop-words, such as 'the', 'and', and 'a' were excluded from the tf-idf score calculations. To calculate tf-idf, we first calculate the number of times each word is used under each topic. The tf-idf score of a word w under a certain topic is equal to the product of the term frequency (tf) and the inverse document frequency (idf). tf is number of times topic contains a word w, and idf is the log of 6 (the number of topics) divided by the number of topics that use word w. The higher the tf-idf score, the more unique word w is to that topic. We were able to gauge the relative engagement with each of these topics and the sentiment towards them. From this analysis, we were able to draw conclusions about the public's response towards the pandemic. These results will be discussed in the next sections.

Results

The following are the results of the tf-idf score calculations of each topics. For each topic, the words with the ten highest tf-idf scores are listed in decreasing order. For definitions of each topic and a discussion on how these topics were decided, please refer to the **Methods** section.

- **public policies**: 'restrictions', 'indoor', 'mask', 'restaurants', 'bonnie', 'citizens', 'dangerous', 'mandates', 'mandate', 'ny'
- **vaccinations**: 'doses', 'arm', 'moderna', 'mrna', 'booster', 'yourself', 'experimental', 'body', 'pfizer', 'vaccinated'
- **variants**: 'omicron', 'variant', 'weigh', 'sentiment', 'fears', 'continue', 'stocks', 'lower', 'milder', 'infectious
- **impact on humans**: 'lebron', 'deaths', 'severe', 'tested', 'james', 'missed', 'crime', 'contagious', 'lot', 'died'
- **political opinions**:'voted', 'vote', 'neglected', 'values', 'jews', 'traitor', 'legislation', 'enable', 'targeted', 'democrats'
- **other**: 'trip', 'communicators', 'product', 'fourseasons', 'province', 'water', 'consensus', 'panel', 'talking', 'jones'

There are a few interesting, and perhaps unexpected, results in the tf-idf calculations. 'lebron' and 'james' appear first and fifth, respectively, under the topic 'impact on humans'. This is likely because around the days that the tweets were collected, the celebrity and basketball star Lebron

James tested positive for COVID, then just a few days later tested negative. His seemingly impossibly fast recovery generated lots of discussion on Twitter. Additionally, the highest scoring words for the 'other' topic do not seem to show a cohesive list of words that center around a common theme, but that is to be expected, since this topic contains miscellaneous tweets that do not fit under any of the other categories. Most of the other high scoring words that categorize the other topics make sense and well represent the what is expected to be discussed under their topics. For example, words with high tf-idf scores include 'doses', 'arm', 'pfizer', and 'moderna'. Pfizer and Moderna are companies who have developed COVID vaccines and to receive a vaccine, doses of it are injected into one's arm. This language is entirely expected in discussion surrounding the vaccine.

Figure 1 showcases the results for topic engagement of the tweets collected. That is, the number of tweets that belong to each topic. The topics with the most engagement are 'vaccination' and 'impact on humans'. These topics were followed by 'other', 'variants', 'public policies', and 'political opinions'. More specifically, the number of tweets corresponding to each topic is: 'impact on humans': 253, 'other': 116, 'political opinions': 86, 'public policies': 90, 'vaccinations': 243, 'variants': 113. 99 tweets were marked 'irrelevant'.

Out of all non-irrelevant tweets, there were 247 negative tweets, 71 positive, and there were 583 neutral. This is shown in **Figure 2**.

Figures 3 to 8 are charts that show the proportions of tweets with positive, neutral, and negative sentiments per topic. With the exception of the topic 'political opinions', the majority of tweets under the topics were neutral. However, most other tweets were coded to be negative. Overall, there were very few positive tweets about the coronavirus pandemic.

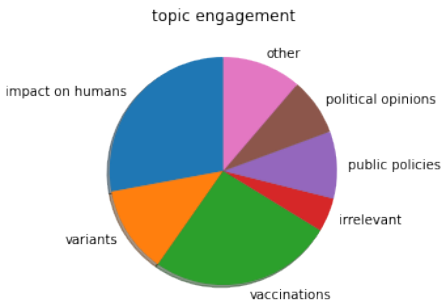

Figure 1: Topic Engagement

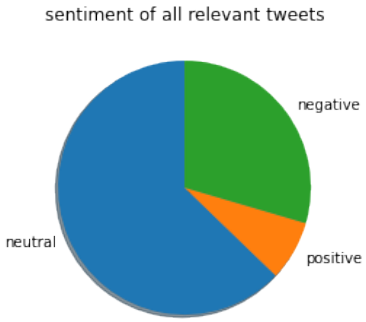

Figure 2: Overall Sentiment (247 negative, 583 neutral, 71 positive)

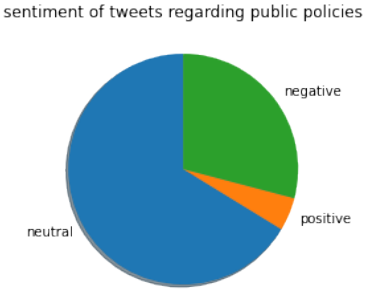

Figure 3: Public Policies Sentiment (27 negative, 5 positive, 58 neutral)

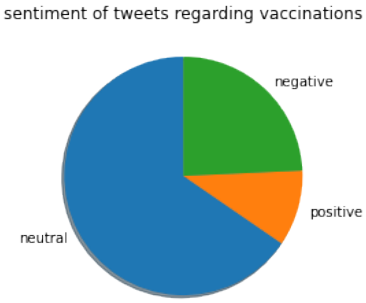

Figure 4: Vaccinations Sentiment (59 negative, 25 positive, 159 neutral)

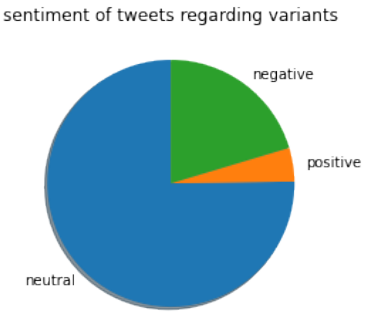

Figure 5: Variants Sentiment (23 negative, 5 positive, 85 neutral)

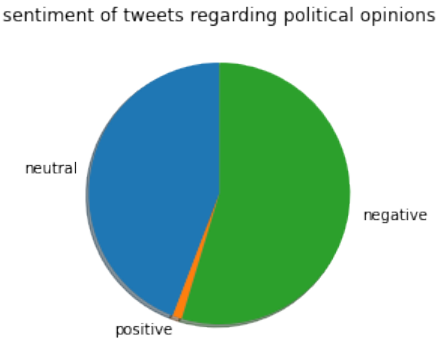

Figure 6: Political Opinions Sentiment (47 negative, 1 positive, 38 neutral)

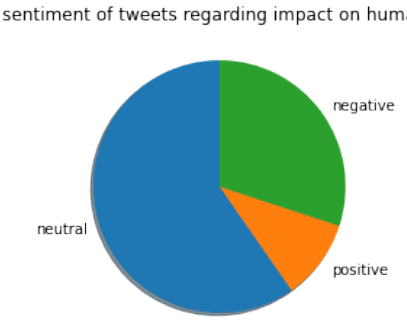

Figure 7: Impact on Humans sentiment (76 negative, 1 positive, 151 neutral)

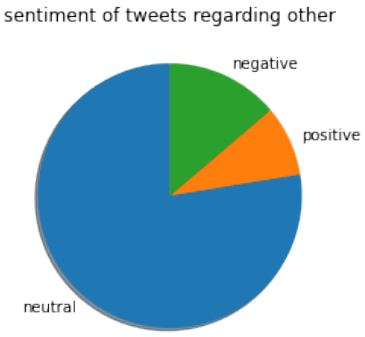

Figure 8: Other Sentiment (16 negative, 10 positive, 90 neutral)

Discussion

The results of the td-idf score calculations and sentiments of different topics show the general concerns with each topic.

Tweets within the 'vaccination' topic held primarily negative or neutral sentiments. Tweets about vaccination with negative engagement were typically about either comments about the dangers of the COVID vaccine, discussions about the experience of receiving the vaccine and/or side effects, or negative remarks to those who refused to receive the vaccine. As reflected in top tf-idf scoring words in the 'vaccination' category, there were many tweets that viewed the vaccine as 'experimental'. We discovered that many tweets using the word 'experimental' suggested the author of the tweet perceived the COVID vaccine to be experimental rather than safe and reliably tested. This indicates there are still a significant amount of people who do not trust the vaccine and refuse to get it because they believe remaining unvaccinated is safer than receiving the vaccine. This perceived danger likely prevents many people from getting a vaccine.

Although there is a great deal of tweets with negative sentiment in the topic of 'vaccinations', it also has one of the highest proportions of tweets with positive sentiments. Other words that appears in the list of top scoring tf-idf words are 'body', 'booster', 'Pfizer', and 'Moderna'. This stems from a number of tweets in which users recount their experience and side effects with the vaccine. We found that many of these tweets were about Twitter users' experiences after receiving a COVID booster shot, which is an additional dose of the COVID vaccine. Discussions regarding booster shots are very salient in this time because they recently been made available to parts of the population in different countries. Some users feel proud after receiving the COVID vaccine, and although may feel some side effects, will tweet about how they are glad to have gotten the vaccine, and encourage others to do the same, leading to more positive tweets. However, negative experiences with COVID vaccines may incite others to be hesitant toward vaccinations. Notably, high tf-idf scoring words included brand names of different COVID vaccines, such as 'Pfizer' and 'Moderna'. This is to be expected, since these companies developed the COVID

vaccines. This shows that Twitter users may be concerned with the effects or efficacy of different types of brands and may compare the effects of them.

Tweets concerning variants mostly center around the new Omicron COVID variant, as reflected in the words with the top tf-idf scores. Most tweets regarding variants were negative or neutral, with less than 5 percent of them being positive. Although many cited concerns with this new variant, the general perception is that this variant shows milder symptoms, as shown with the word 'milder' in the list of top words. This perception may fuel vaccine hesitancy and concerns about whether the current vaccines will protect people from these new variants. There are also people who believe that because the variants of COVID appear to be quite mild, a vaccine isn't necessary. However, high scoring tf-idf words such as 'fears' and 'infectious' show that there is a significant portion of the public that is concerned about the new coronavirus variants that appear.

Next, we will consider tweets under the 'public policies' topic. Tweets under this topic were about new 'restrictions' and 'mandates', as shown by the list of high scoring tf-idf words. Most tweets carried a neutral sentiment. For example, many tweets came from News Twitter accounts or other users who simply stated any newly enacted restrictions, but did not necessarily have an opinion or particular sentiment about those new restrictions. The restrictions and mandates that were of most concern amongst the Twitter community were restrictions in 'restaurants' and 'indoor' spaces, as reflected in the tf-idf word list. Although many tweets were simply stating new restrictions due to COVID and carried a neutral tone, about one third of the tweets carried some sort of sentiment. The majority of those tweets were negative. We found that some Twitter users felt that the restrictions were too harsh, while others felt that the restrictions were not enough to curb the dangerous spread of COVID.

The topic with the most engagement is 'impact on humans' and it is clear why. The coronavirus pandemic has impacted essentially every humans' way of living and has taken the lives of countless loved ones. Words such as 'severe', 'died', 'deaths', and 'contagious' appeared in the list of high scoring tf-idf words. These words indicate that Twitter users tweeting about the human impacts of the coronavirus pandemic are serious and many people's lives have been lost due to this devastating disease. Although the impact is largely negative, the majority of tweets under this topic were categorized as having neutral sentiment. This is likely because many tweets, for example, tweets from News Twitter accounts, would state the number of deaths and positive COVID cases in a particular region. Additionally, some accounts tweeted about certain celebrities who contracted COVID. Many of these tweets did not use any strong positive or negative language. However, if a tweet was coded to have a non-neutral sentiment, it was usually coded to be negative. Many users used Twitter to express their frustrations with the loss of their loved ones or talked about fears they have for their friends and relatives who have contracted the disease. Words such as 'missed' and 'severe' that appear in the tf-idf word list show that Twitter users missed those who have passed, and may had connections with people who suffered very severe symptoms of the virus. It's clear many Twitter users are upset with the high death toll and impacts that COVID has had in their lives and the lives of others effected.

Surprisingly, the topic 'impact on humans' also had one of the greatest ratio of positive tweets to non-positive tweets. Many users tweeted about the good deeds that members of their communities have been doing to help those who were negatively affected by the pandemic, such as those from low-income households. There was also a number of tweets commending essential service workers, whom continued to work despite the dangers of the pandemic. Although the coronavirus pandemic has brought on lots of troubles and pain, the tweets gathered show that the human psyche can still stay positive and support other humans during these tough times.

The most revealing topic regarding vaccine hesitancy may be the topic of 'political opinions'. The word 'value' appears to be very unique to this topic, as seen in its high tf-idf score. This indicates that the political views Twitter users' have in response to politicians responses to the coronavirus pandemic are largely reflective of their personal ideologies. Many words with negative connotations appear in this topic's list of highest tf-idf words such as 'neglected' and 'traitor'. This shows that many Twitter users believe that politicians have been handling the coronavirus pandemic very poorly, and in a way that violates their values. Many tweets criticize the politicians and members of certain political parties who support certain restrictions and mandates because it infringes on their values relating to freedom. Other users criticize politicians and members of certain political parties for not abiding to established restrictions or accusing them of not caring about those who are severely effected by the pandemic. It is probable that those who hold the perception that their value of freedom is being restricted are more likely to hesitate to get a vaccine. Those who feel as if they have no freedom would not appreciate being told that COVID vaccinations were are a requirement and may rebel by refusing to get the vaccine. Overall, the majority of the tweets regarding political opinions have negative sentiments. No matter the Twitter users' political leanings, their views on the politics relating to the pandemic have been very negative.

Data analysis on the 'other' category does not reveal much about Twitter users' views on the coronavirus pandemic. This is primarily because the tweets in this category mentioned the pandemic, but gave no insight on the Twitter users' views on the pandemic and did not fall into any of our defined topic categories. For example, one tweet marked 'other' simply said the word 'covid' and contained no other text. Another example is of a tweet linked to a humorous picture that the Twitter user found 'at the beginning of the pandemic', but the contents of the tweet say nothing the users' views on COVID. It seems that the pandemic has become a part of everyone's daily life in such a way that many tweets about various topics contain key words pertaining to the pandemic. However, these tweets do not give insight on the users' thoughts on the pandemic itself.

Overall, the topics that provide the most insight on vaccine hesitancy are 'political opinions' and 'vaccinations'. Within those topics, we see that many of the negative sentiments stem from those who view vaccines as experimental, unreliable, and unsafe. We also see many Twitter users accuse politicians of ignoring their values of freedom by instating many restrictions. These perceptions of an unsafe vaccine and violation of values all seem to fuel vaccine hesitancy and a negative response to the pandemic.

Contributions

Yixuan Li wrote the program for gathering tweets and calculating the td-idf scores. She also created the charts used in the Results section.

Darien Hong wrote the Abstract, Introduction, Data, and Methods sections of the report.

Grace Qian wrote the Results, Discussion, and Contributions sections of the report.

All three group members worked together to open code tweets to create 6 topics for discussion. The 1,000 tweets gathered were split amongst the group members to annotate. Furthermore, all group members helped give feedback and made edits on other group members' parts of the report