COMP 598 — Final Data Science Project: COVID in Canada

Written by Nicolas Fertout, Ben Zhang, Aiden Cho.

nicolas.fertout@mail.mcgill.ca
jaeshin.cho@mail.mcgill.ca
yuteng.zhang2@mail.mcgill.ca

Introduction

The last two years have been severely impacted by the global COVID pandemic. With the recent vaccine policies, social media has become a true battleground where people express their ideas, argue and communicate their feelings.

In this report, we investigated the discussions currently happening on Twitter around Covid-19 in Canada, and especially regarding vaccine hesitancy. Our team collected opinions from tweets within 3 consecutive days, and analyzed the overall trends with regard to Covid-19 and vaccination. Our focus was on identifying the main topics discussed along with their corresponding engagement, as well as the overall response to the pandemic and vaccination after almost 2 years. Using the discussions surrounding the topics Vaccination incitation, Vaccination dissuasion, Symptoms, Vaccine side effects, Policies, Repercussions, News and Jokes, we managed to identify the main trends and behaviors, while doing topic-specific analyses.

Overall, using those topics, in addition to a sentiment classification and a tf-idf weighting factor, we found that the Covid-19 crisis is portrayed as a true trauma, at the Canadian and international scale. We also discovered that a lot of agitation surrounds the vaccination process, with a large proportion of tweets involving feelings and persuasion methods, showing that vaccine hesitancy is a salient part of discussions around Covid. More generally, it seems that social media users (and in this case twitter) tend to express their disagreement and indignation in an unusually strong and pronounced manner.

We will start by explaining what our data consists of and how it was collected, then we will give some insights on the methods and design choices made during our project, to finish with a presentation of our results and an interpretation of our findings.

Data

Our data consist of english covid-related tweets collected within a span of 3 days. We built a twitter API to scrape data (using the `tweepy` library on python), and gathered 3000 raw tweets (1000 tweets per day). As part of our query, we used keywords “covid”, “vaccine”, “vaccination”, along with some of the most popular vaccine brand names in Canada: “Pfizer”, “Moderna”, “Johnson and/& Johnson” and “AstraZeneca”. We then applied some automated filtering to our data, selecting tweets in english only, and removing duplicates and retweets of original tweets already present. Finally, we removed additional irrelevant tweets by hand, and selected at random 1000 tweets from the remaining.

Table 1 summarizes the number of tweets obtained after some automated filtering, and then after some human filtering.

	Day 1	Day 2	Day 3	Total
Raw data collected	1000	1000	1000	3000
After automated filtering	614	576	702	1892
After human filtering and random selection	340	304	356	1000

Table 1: Number of tweets obtained after each step

We tried to get around the same number of tweets for each day, so that the overall opinion is not affected by an event occuring on a specific day.

The dataset inclusion of the keywords used is summarized in Table 2.

Keyword	Proportion of tweets
covid	45.7%
vaccine	68.6%
vaccination	11.9%
pfizer	6.4%
moderna	4.7%
johnson & johnson	1.8%
astrazeneca	2.9%

Table 2: Number and proportion of keywords in the data

As one could expect, the majority of the tweets collected contains the word “vaccine” or “covid”, with some tweets referring to specific brand names.

Methods

Collecting, annotating and analysing the data required some design decisions, in order to get the most accurate results possible.

Data Collection

A lot of government policies are announced every day, and almost every hour. Since tweets about covid are so numerous, we could easily collect 1000 of those within 1 hour. However, we did not want the opinions to be biased or directed towards a specific news that came out at this time. So we collected 200 tweets, at 5 different time slots, for three consecutive days. This helped dealing with unbalanced data. Some spammed tweets, along with some tweets referring to vaccination to other diseases were not detected by our filtering algorithm, so we had to do some additional filtering by hand.

Data Annotation

When having to choose the topics, we first manually went over 200 tweets to grasp what the tweets collected are usually about, and to identify the different categories we could get. During this open-coding phase, we tried to get a full partition of the dataset of tweets, so that we would not need an "other" category. The selected topics are defined in detail in the Result section. Besides of the topic annotation, we also assigned each tweet with a sentiment: either positive, neutral or negative. This additional variable is meant to add information on the engagement and involvement of users, both globally and per topic.

Data Analysis

Finally, during the analysis, we focused on engagement in every topic, feelings for each topic, as well as a list of the most used words per topic, using the tf-idf method. In the latter, we chose to remove common English words from those lists (by using stopwords), along with the keywords "covid", "vaccine" and "vaccination", in order to get a more meaningful and accurate characterization.

Results

Our first analysis step was the choice of topics. We wanted to choose salient topics discussed around covid-19 that partition the whole dataset. During the open-coding phase, we had the opportunity to grasp what the tweets are generally about, the subjects discussed, as well as who wrote them (individuals, meme pages, news media, ...). We tried our best to display the overall trends and engagements, while being specific enough to characterize them and give further insights and understandings.

Here are the 8 topics chosen and their description, along with one example of a tweet in this topic from the collected dataset:

1. Vaccination incitation

   Description: tweets where individuals incite or encourage vaccination.

   Example: "Booster shots are now available for many New Yorkers who received the Pfizer, Moderna, or Johnson & Johnson #COVID19 vaccines!Go get your shots!"

2. Vaccination dissuasion

   Description: tweets where individuals dissuade or discourage vaccination.

   Example: "hey remember how they were vaccinating an entire city in brazil a couple months ago? why haven't we heard about the amazing success of that city in ending the pandemic for their residents? or how safe it was? ????"

3. Policies

   Description: tweets discussing government restrictions, regulations policies, control measures and health facilities.

   Example: "Are governments going to set new restrictions in the UK due to the high numbers of covid case?Seriously!..."

4. Repercussions

   Description: tweets discussing socioeconomic changes due to the pandemic and vaccination (and not directly related to government policies).

   Example: "For months, medical experts warned that leaving large areas of the world unvaccinated would make new variants inevitable."

5. Symptoms

   Description: tweets where individuals speak of their covid-19 symptoms.

   Example: "Update: My fever ?? finally broke this morning and I'm feeling much improved from yesterday. I'm still sore"

6. Vaccine side effects

   Description: tweets where individuals speak of their vaccine side effects.

   Example: "A quick question I had my booster Monday (Moderna) I had Moderna as my 2 jabs too all ?? sore ?? that was it. Today I'm in so much pain (muscle) can't pick anything up and it hurts walking up the stairs, Ifeel I've ran a marathon as pain is only bottom half. anyone else had this?"

7. News

   Description: tweets by news media, article links, usually about the overall pandemic, and not written by an individual.

   Example: "Gravitas: Revealed: How Pfizer blackmails countries for shots https://t.co/latIeXrgZQ"

8. Jokes

   Description: memes and jokes about covid and vaccination.

   Example: "Someone in this store literally just gave a covid fart because we could smell that shit through our masks"

Once the annotation is done, we can start analyzing our data. We provide below tables and chart to better understand the topics distribution, engagement and characterization. Results are reported in this section, and interpreted is the Discussion section.

Topic Distribution

Figure 1 is a pie chart representing the distribution of topics. Percentages are reported in Table 3.

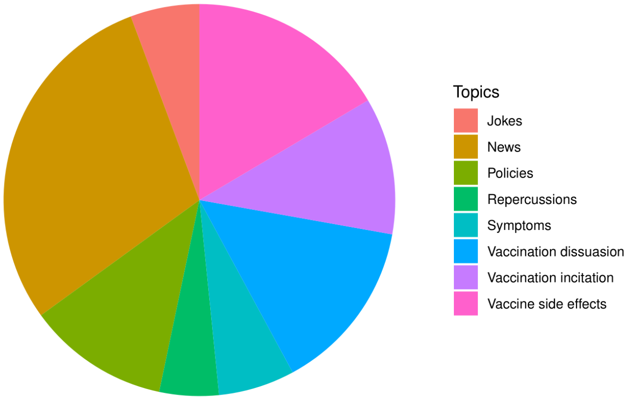

Figure 1: Pie Chart of the Distribution of Topics

Topic Name	Proportion of tweets
Vaccination incitation	11.3%
Vaccination dissuasion	14.3%
Policies	11.7%
Repercussions	4.9%
Symptoms	6.3%
Vaccine side effects	16.5%
News	29.3%
Jokes	5.7%

Table 3: Topic Distribution

As we can see from Table 3 and Figure 1, the dataset consists mainly of tweets written by news media (29.3%), which is expected considering how sensitive this subject is in Canada and around the world. A lot of tweets about the vaccination were also collected, with more than 400 tweets either discussing their side effects (16.5%), or encouraging/discouraging others to get vaccinated for covid (11.3%/14.3% respectively). Besides, around a quarter of the tweets address general issues around covid and vaccination, such as government policies (11.7%), socioeconomic changes (4.9%), and covid symptoms (6.3%), which shows how much the disease and measures affect people, even after 2 years of ongoing global crisis.

Topic Engagement

Figure 2 (see top of next page) is a stacked bar chart representing the proportion of sentiment per topic. Numbers and proportions are reported in Table 4.

Topic Name	Sentiment	Number of tweets	Proportion per topic
Vaccination incitation	Positive	60	53.1%
	Neutral	44	38.9%
	Negative	9	8.0%
Vaccination dissuasion	Positive	12	8.4%
	Neutral	75	52.4%
	Negative	56	39.2%
Policies	Positive	5	4.3%
	Neutral	46	39.3%
	Negative	66	56.4%
Repercussions	Positive	3	6.1%
	Neutral	19	38.8%
	Negative	27	55.1%
Symptoms	Positive	10	15.9%
	Neutral	32	50.8%
	Negative	21	33.3%
Vaccine side effects	Positive	19	11.5%
	Neutral	33	20.0%
	Negative	113	68.5%
News	Positive	29	9.9%
	Neutral	227	77.5%
	Negative	37	12.6%
Jokes	Positive	2	3.5%
	Neutral	30	52.6%
	Negative	25	43.9%
Total	Positive	140	–
	Neutral	506	–
	Negative	354	–

Table 4: Sentiment Distribution per Topic

This sentiment-per-topic analysis highlights many interesting features. First, as one could expect, the total number of tweets per sentiment reveals the trauma of the crisis on social media users (140 positive against 354 negative – Table 4). When removing the "News" to only consider tweets written by individuals, we get 111 positive, 279 neutral and 317 negative sentiment (respectively 16%, 39% and 45% of tweets written by individuals). Almost half of those tweets display negative feelings about covid and vaccination.

This trauma is reinforced when we take a look at the sentiment distribution per topic. While News are mostly neutral (77.5% of neutral sentiment), other topics seem more divided. Some of them have a majority of negative sentiment (Vaccine side effects, Repercussions and Policies, with 68.5%, 55.1% and 56.4% resp.), while other are more uniformly split (like Symptoms). An interesting parallel can be made when observing Vaccination dissuasion/incitation topics. The dissuasion ones are around 8% positive and 40% negative, while the incitation ones are 53% positive and 8% negative. This negative matching will be discussed in the Discussion section.

Topic Characterization – tf-idf

Table 5 reports the 10 words with the highest tf-idf score per category.

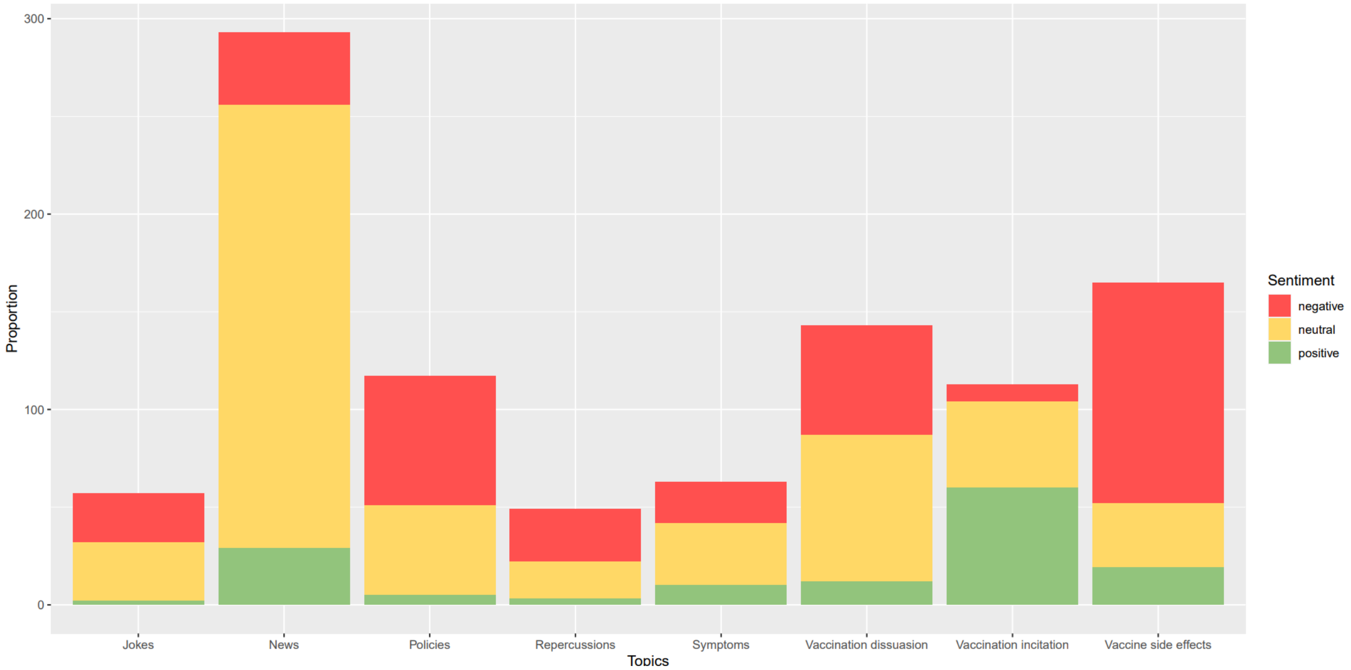

Figure 2: Stacked Bar Chart with Sentiment per Topic

Topic	10 highest scored words
Vaccination incitation	vaccinated, dose, success, omicron, immunity, effective, pfizer, booster, shots, delta
Vaccination dissuasion	stop, omicron, interchangable, variant, administration, evidence, children, emergency, punishing, eua
Policies	dose, available, health, clinic, moderna, children, eligible, booster, doses, two
News	pfizer, people, countries, federal, judge, reject, omicron, variant, moderna, two
Repercussions	world, variant, countries, companies, clean, pandemic, experts, profits, threat, africa
Symptoms	rejects, health, pharmaceutical, pharma, head, anitbody, fiver, better, home, sick
Vaccine side effects	sore, pain, hurts, dies, pfizer, plumber, getting, yesterday, effect, moderna
Jokes	twitter, love, lord, moderna, omicron, lol, huge, haha, experts, sneezes

Table 5: 10 Highest Scored Words using tf-idf Analysis per Topic

The tf-idf is a weighting statistic which increases proportionally to the number of occurrences of a word in a given category, but decreases proportionally to the number of occurrences of that same word in all other categories.
Those words will be used in our interpretation in the next section. Overall, some words are present in several topics, like ”doses” or ”omicron”, while others are topic-specific and give a good understanding of the characterization of each category (for instance ”dies” is Vaccine side effects, or ”world” in Repercussions).

Discussion

A Canadian and International Crisis

Analyzing Canadian tweets shed light on the importance of the Covid crisis on different levels. Indeed, a large portions of the tweets are in the News category and almost 15% of them talk about policies and repercussions. From the tf-idf result, we notice that the word “federal” appears very frequently in the News topic highlighting the large amount of medias referring to Covid in Canada.

However, the crisis is also international. The 3 most frequent words in the Repercussions category are “world”, “variant” and “countries”, showing how Canadian news and people also address concerns about international matters in relation with the Covid-19. The word “Africa” also shows up a lot in the repercussions-related tweets which accentuate this international perspective discussed above.

It also opens the discussion on the different variants that have world-widely worsen the Covid-19 situation. This analysis demonstrates that variants are a noteworthy discussed matter since a variant-related word appear in 4/8 topics (omicron and delta for the “vaccination incitation” topic, omicron and variant for the “vaccination dissuasion” topic, variant for the “news” topic, and omicron in “jokes”). The newest African variant (Omicron) have alerted many countries such as the UK which has put in place stricter regulations. This finding raises the fact that variants have prompted considerable panic across the globe.

Taking the vaccine : a salient question

We also focused our analysis on vaccine hesitancy and response. The main reason vaccines are part of our lives is because of the severity of Covid-19. The symptoms of this disease being very harmful, governments and pharmaceutical firms had to come up with a common solution to fight this virus. We observe that 6.3% of the gathered tweets are about symptoms which occured when catching the virus. Even if most of those tweets are neutral, there is a larger proportion of negative sentiment regarding the symptoms : 33.3% against 15.9% for positive. The tf-idf strengthens this idea since 6 words out of 10 express pain-related feelings (“rejects”, “health”, “head”, “antibody”, “fever”, “sick”).

Consequently, to reduce the violence of Covid-19 or at least to ease the current symptoms, many people promoted vaccination. Our search reveals that 11.3% of the tweets collected encourage people to get vaccinated. The 10 highest scored words of this category contain “success”, “immunity“, “effective” and “vaccinated”. It seems like vaccinated people praise the benefits of the vaccine and confirm its effectiveness. Moreover, we note that the word “better” appears in the tf-idf of the Symptom topic, confirming people are healing.

Finally, one of the most surprising results we came across is that the “Vaccine side effect” topic covers 16.5% of the collected tweets. This topic has the largest proportion of negative sentiment (68.5%) compared to the mean of all the topics (39%). The tf-idf table for this topic contains very strong negative words such as “sore”, “pain”, “hurts” and “dies” showing the bad response to the vaccine the canadian society experienced and related.

Hence, we can conclude that people had very different experiences of symptoms, vaccination and side effects since the spectrum of our result is very broad.

Twitter : a media to express disagreement

A really surprising revelation is that, even though 80.13% of Canadians received at least one dose of a vaccine, there are still more tweets that dissuade people from taking the vaccine than tweets that encourage them. We also know that people tend to find the vaccine effective (see previous part) which makes the result even more surprising.

For instance, even if the joke category is not the most pertinent, it reinforces the overall trend of disturbance regarding the pandemic, as most jokes and memes criticize policies, experts or the vaccination process (from the tf-idf analysis and sentiment distribution) . People are fed up, and they use humour to express it. Moreover, the third most discussed topic is Vaccination dissuasion (14.3% of tweets collected) with 39.2% of negative sentiment. The words from this category are very strong (i.e “stop”, “punishing”, “evidence”) revealing that some people strongly disagree with taking the vaccine.

Although these results seem very strong, it also highlights that social media tends to be more used to express disagreement and indignation. Apart for Vaccine incitation, all the different topics have more negative sentiments than positive ones which fortifies our point. We can see a strong association: individuals that tend to discourage use negative sentiments, usually with anger and irritability, while individuals encouraging vaccination will use more tact and be more expressive about the positive outcomes that can be obtained from getting vaccinated. In both cases, we observe that people use pathos as their communication technique to persuade others, which is common in social media, especially around sensitive subjects like Covid or vaccination.

Group Member Contributions

- Everyone
  - Global planning of the project
  - Recurrent zoom meeting to discuss each member’s advancements and define next steps
  - Conduction of the open coding phase to define the topics
  - Manual filtering of the data
- Aiden Cho
  - Annotation of the tweets by defining for each tweet a sentiment and a topic
  - Writing of a script to perform the tf-idf analysis
- YuTeng Zhang
  - Creation of a python script and an API to collect and filter the tweets
- Nicolas Fertout
  - Data analysis and result interpretation
  - Creation of figures and tables for the different section
  - Writing, editing and formatting of the final report

References

Le Monde et AFP. 2021. Variant Omicron : un raz-de-marée arrive au Royaume-Uni, Boris Johnson prend une série de nouvelles mesures. https://www.lemonde.fr/international/article/2021/12/12/face-a-la-progression-du-variant-omicron-le-niveau-d-alerte-covid-releve-au-royaume-uni_6105791_3210.html

Canada, P. H. A. o. 2021. Demographics: COVID-19 vaccination coverage in Canada. https://health-infobase.canada.ca/covid-19/vaccination-coverage/

Thank you for reading this report!