COMP 598 Final Project - COVID-19 in Canada

Xinyi Zhu[1*], **Leopoldo Zugasti** [2*], **Svetlana Simantov** [3*]

3480 Rue University, Montréal, QC H3A 2A7
[1] xinyi.zhu3@mail.mcgill.ca
[2] leopoldo.zugasti@mail.mcgill.ca
[3] svetlana.simantov@mail.mcgill.ca

Introduction

In this project, our objective was to understand recent discussions around COVID-19 in Canadian social media. In particular, we were interested in the salient topics about COVID-19 and what each topic primarily concerns, relative engagement with these topics, and the sentiment in response to the pandemic or vaccination.

Our analysis drew on Twitter posts, or tweets. We collected 1,000 unique English tweets within a three-day window, filtering by the contents that mentioned certain keywords such as "vaccine" and "pfizer". We conducted an open coding on 200 tweets and developed eight topics, namely "mandate", "efficacy", "unvax", "political", "side_effect", "information", "children", and "spread". We manually annotated each tweet with one of these topics, as well as with a positive, negative, or neutral sentiment.

With fully annotated 1,000 tweets in our dataset, we computed the ten words in each topic with the highest TF-IDF scores, and conducted further analysis to gain better insight on the Twitter climate surrounding COVID-19 and pandemic measures.

The amount of positive and negative sentiments across all tweets in the dataset are balanced with a slight skew for the positive. The positive sentiments mainly come from the categories "unvax" and "spread". However, the top two categories with the most relative engagement are "mandate" and "efficacy" are mainly leaning towards negative sentiments rather than positive. Although there is more of a positive sentiment across the entire dataset, Twitter users ultimately interact with categories that have more negative attitudes, thus promoting vaccine hesitancy on the social media platform nonetheless.

Data

The dataset used for this project consists of 1,000 English tweets containing any of the following keywords within the text of the post (including hashtags): 'covid', 'vaccination', 'vaccine', 'pfizer', 'BioNTech', 'j&j', 'johnson&johnson', 'moderna', 'astrazeneca' and 'astra zeneca'.

Moreover, these tweets were posted by Canadian users. A 'Canadian user' is anyone that manually set their location to 'Canada' or a Canadian province, or had their geolocation enabled on Twitter and the country code associated with their location was 'CA'. We are aware that this was removed as a requirement but we kept it since we had already collected our dataset by the time of the announcement.

Further, we manually removed any tweet that was unrelated to COVID-19. Some examples include tweets that had the keyword 'vaccine' or 'vaccination' that were related to a non-COVID-19 vaccine, or tweets that mentioned brand names that were related to other products these companies develop. Note that these instances occurred very rarely; roughly 0.7% of originally collected tweets were manually removed.

Originally, we collected 3,000 tweets between November 12, 2021 and November 14, 2021, inclusive. Between these three inclusive days, 1,000 tweets were collected, where we then randomly sampled 1,000 of the 3,000 total tweets collected to create our dataset. All manually removed tweets were replaced by randomly sampling the remaining 2000 tweets. The random sample gave us 363 tweets from November 12, 306 from November 13 and 331 from November 14.

During the three-day window, it's important to note that Pfizer was in the process of approving its vaccine administration to children aged 5 to 11 years old (Health-Canada 2021). Since Twitter users discussed this new development extensively in relation to overall COVID-19 conversation, the tweets collected and included in our final dataset strongly reflects this topic.

Methods

Experimental Setup

In order to obtain 'Canadian users', we used an imperfect solution due to technical limitations. Indeed, the Twitter API rarely provides geographical information mostly due to the fact that users don't usually choose to provide their geolocation. Therefore, we approached the problem in a different manner: we used a 'location' attribute that users can set manually to any value. If this value contained the word 'Canada' or a name of a Canadian province, we assumed that this tweet was of Canadian origin. This technique greatly improved the amount of tweets we were able to retrieve. However, it's significant to be aware that because this value is

[*] These authors contributed equally.

set manually by Twitter users, we cannot guarantee that this information is accurate. Nevertheless, we were able to confirm qualitatively that this heuristic worked very well when manually annotating.

In addition to the keywords suggested in the instructions, we tried to include multiple different ways to refer to each branch in order to cover the topic with as much generality as possible. For example, BioNTech and Pfizer refer to the same brand, thereby both words were included as keywords. Although not specified in the instructions, we also included the keyword "vaccine" because this word would permit us to cover a larger range of COVID-19 vaccine-related tweets.

Data Collection

Instead of collecting 333 tweets each day directly, we decided to collect 1,000 tweets each day and then sample 1,000 out of the 3,000 total tweets mostly due to technical reasons. For example, it was easier to resample the data to replace any manually deleted tweets during the open coding and manual annotations phases.

During the data collection phase, we decided to collect replies as well as original posts. Although we didn't include retweets — as they include the same content as the original posts collected — replies were collected because they functionally fulfill the same role as the retweets as they tend to agree with the original tweet. Replies, in general, fuel and extend the conversation started by the original tweet, providing a better insight into the topics discussed and the sentiments felt by the Twitter usership.

Regarding the manual removal of tweets, we discussed the removal of a specific tweet as a group in order to remain as impartial as possible. There must be a unanimous decision between all members of the team to remove that tweet if it was not entirely related to COVID-19.

Data Annotation

The data annotation was started with an open coding of the first 200 tweets collected. Each member of the team annotated these independently. We then proceed to go over all of them in a group and compare our categories/annotations and come to a consensus. Once a clear definition of each category was properly defined, the remaining 800 tweets were split in three equal parts and annotated individually by each member of the team.

TF-IDF Scores Computation

When computing the top 10 words with the highest TF-IDF scores for each category, we wanted to filter out certain characters and punctuations. For all 1,000 tweets in our dataset, we removed the stopwords, replaced certain punctuations with a space, removed any non-alphanumeric words, made sure all words collected were case insensitive, and only kept words with a frequency with at least 5. Our TF score formula computes the number of times a specific category has the specified word. Meanwhile, the IDF and TF-IDF score formulas are as follows.

$$TF(w, t_c) = freq(w, t_c)$$

$$IDF(c, c_w) = \log \frac{count(c)}{count(c_w)}$$

$$TF\text{-}IDF = TF(w, t_c) * IDF(c, c_w)$$

where w is the word, t is the tweet, c is the category, t_c is all the tweets of a specific category (ie. a category document), c_w is the category that uses the specified word.

Results

Definitions

After conducting the open coding of 200 tweets, we selected eight topics. The following are their names and definitions:

Mandate: An official order/requirement that states one must follow a mandate, such as being vaccinated, attaining a vaccine passport, wearing a mask indoors, etc. in order to do certain activities, including working, travelling, going to restaurants, etc.

Efficacy: The effectiveness — or lack thereof — of vaccines, mandates, masks, and any other measure in curbing the pandemic

Unvax: Discussion about unvaccinated/anti-vaccine people

Political: Criticism or general discussion of the government or a political figure

Side_effect: Discussion of side effects pertaining to any COVID-19 measures/vaccines

Information: Either a useful statement (eg. "Here are the vaccination centers."), or a quest to obtain information (eg. "Where can I get vaccinated?")

Children: Anything related to children, their activities, and overall relation to COVID-19 and pandemic measures (eg. vaccination approval for children)

Spread: The transmission of COVID-19 or how to prevent it (eg. contact tracing)

Note that **positive**, **negative** or **neutral** attitudes are towards vaccines, mandates, and other pandemic measures and not attributed to tone. For example, a tweet may say "I despise unvaccinated people!". While the tone and word choice suggests a negative sentiment, we would annotate this hypothetical tweet as having a positive sentiment because their attitude and position demonstrates that they are pro-vaccine.

Topic Engagement

We have the relative topic engagements for each category, as well as the breakdown of sentiments across the entire dataset, expressed in Figures 1 and 2, respectively. In descending order, the topic engagement is as follows: "mandate" (22.4%), "efficacy" (21.9%), "unvax" (13%), "political" (11.5%), "side_effect" (11.3%), "information" (9.3%), "children" (5.4%), and "spread" (5.1%).

Sentiment Breakdown

In addition to computing the top 10 words with the highest TF-IDF scores for each topic, we also produced the sentiment breakdown for each category. With these results in mind, we saw that the overall positive sentiments across the entire dataset originated mainly from the "unvax" (73.1% for

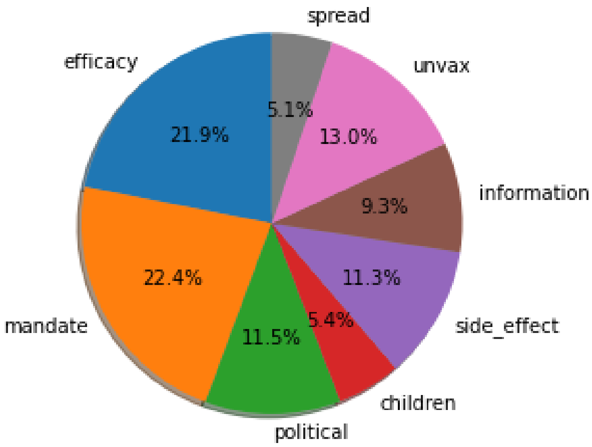


Figure 1: The relative topic engagements for each category

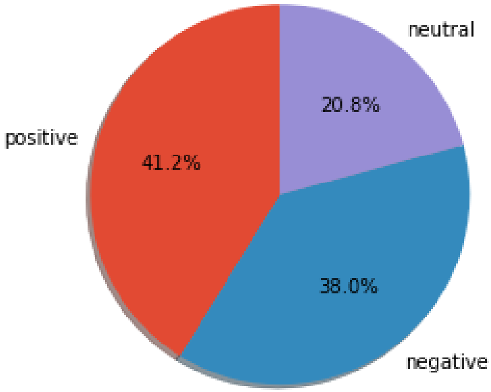


Figure 2: The breakdown of sentiments

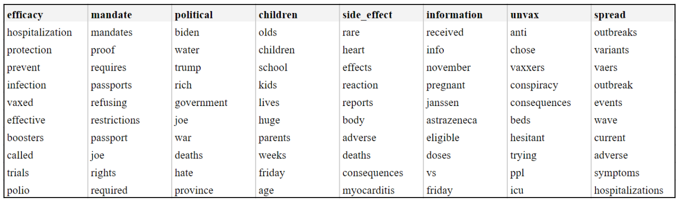
positive sentiments) and "spread" (51% for positive sentiments) categories. Yet the "mandate", "efficacy", "political", and "side_effect" categories skewed more towards the negative sentiments by 49.1%, 47.9%, 42.6%, and 53.1%, respectively. The "children" category has the same percentage for both positive and negative sentiments at 37%. Out of all the categories, "information" has an overwhelming amount of neutral sentiments at 80.6%.

Topic Characterization

A comprehensive list of all the top 10 words with the highest TF-IDF scores for each category, as well as the sentiment distribution for each category, are shown in Figures 3 and 4. The scores are displayed from high to low from top down.

Discussion

As we were collecting, annotating, and analyzing our dataset, we were surprised to see no huge polarity between the amount of positive and negative sentiments across all tweets in the dataset. Overall, we have 41.2% positive sentiments, 38% negative sentiments, and 20.8% neutral sentiments out of 1,000 total annotated tweets. Not only are the positive and negative sentiments around the same amount, but surprisingly there are slightly more positive sentiments when it comes to vaccine hesitancy and overall attitude towards the pandemic. At first glance, this would mean that there are more Twitter users that are pro-vaccine, pro-mandate, and are more informative and safe about the pandemic.



efficacy	mandate	political	children	side_effect	information	unvax	spread
hospitalization	mandates	biden	olds	rare	received	anti	outbreaks
protection	proof	water	children	heart	info	chose	variants
prevent	requires	trump	school	effects	november	vaxxers	vaers
infection	passports	rich	kids	reaction	pregnant	conspiracy	outbreak
vaxed	refusing	government	lives	reports	janssen	consequences	events
effective	restrictions	joe	huge	body	astrazeneca	beds	wave
boosters	passport	war	parents	adverse	eligible	hesitant	current
called	joe	deaths	weeks	deaths	doses	trying	adverse
trials	rights	hate	friday	consequences	vs	ppl	symptoms
polio	required	province	age	myocarditis	friday	icu	hospitalizations

Figure 3: Top 10 words with the highest TF-IDF scores for each category

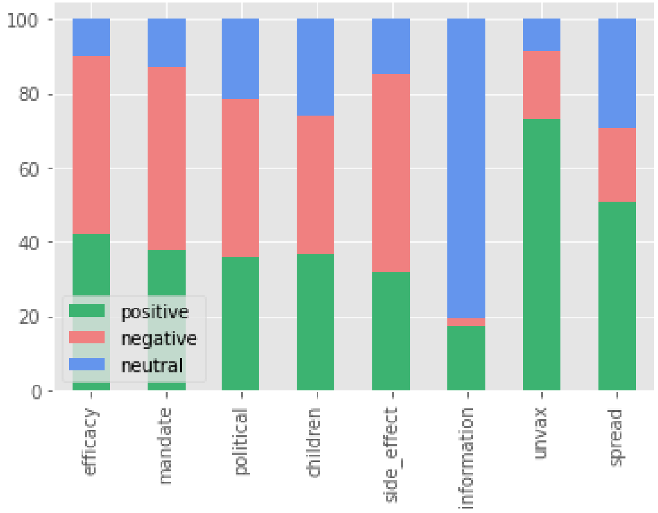


Figure 4: The sentiment distribution for each category

Yet as we look into the percentage of sentiments per category, we can still see that Twitter is more concentrated with negative sentiments due to its usership's relative engagement with specific categories.

Let's look at the level of engagement for every category across our entire dataset, ordered from most to least engagement: "mandate" (22.4%), "efficacy" (21.9%), "unvax" (13%), "political" (11.5%), "side_effect" (11.3%), "information" (9.3%), "children" (5.4%), and "spread" (5.1%). Taking the top two categories with the most relative engagement among the Twitter usership — "mandate" and "efficacy" — we can look at their sentiment distribution to see if these categories are skewing more towards positive or negative attitudes towards vaccines, mandates, and the other pandemic measures. Both categories are skewing more towards negative sentiments than positive: "efficacy" has 47.9% of negative sentiments compared to 42% of positive sentiments, and "mandate" has 49.1% of negative sentiments versus 37.9%

of positive sentiments. Since both categories have more negative sentiments, their top 10 words with the highest TF-IDF scores have a more negative connotation. For example, "efficacy" has popular words such as "protection", "effective", and "prevent". With the overwhelming percentage of negative sentiments in this category in mind, these rare words are suggesting that Twitter users are most probably questioning the efficacy, protection, and prevention of COVID-19 when following pandemic measures (vaccines, mandates, masks, etc.). The same mentality applies to the category "mandate". Some of this category's words with highest TF-IDF scores are "rights", "proof", and "refusing", demonstrating Twitter users' mandate refusal and hesitancy, including vaccines. Having the two most engaging categories leaning towards more negative sentiments ultimately dilutes the conversation on Twitter regarding vaccine hesitancy, COVID-19, and the pandemic overall. Since Twitter users are publishing posts and engaging with tweets that mainly fall under the categories of "efficacy" and "mandate", the sentiment on vaccine hesitancy is overwhelmingly negative across the entire dataset.

But we still have to account for the fact that the overall sentiment analysis still skews slightly for the positive: where are those positive sentiments coming from? The overwhelming positive sentiments seem to predominantly stem from the following two categories: "unvax" and "spread" with positive sentiments of 73.1% and 51%, respectively. Since "unvax" is the third most engaged category, its positive sentiments is most likely responsible for the total percentage of positive sentiments across the entire dataset. However, the tweets that fall under the "unvax" category, although mostly express positive sentiment, still reference negative attitudes of others in regards to the pandemic and its measures. We can see that with the top 10 words in the "unvax" category with the highest TF-IDF scores, including "anti", "vaxxers", "conspiracy", and "hesitant". Despite containing the majority of positive sentiments in the dataset, the "unvax" category still references words and concepts that pertain to those upholding negative attitudes, thereby emphasizing the impact of those negative sentiments rather than the positive sentiments expressed by the publisher of the tweet itself. The "spread" category doesn't help either because regardless of its overwhelming amount of posts with positive sentiments, the category itself is the least interacted and unpopular with the Twitter usership out of all eight categories; in other words, the "spread" category barely makes an impact on the overall percentage of positive sentiments in our dataset and perhaps the entire Twitter conversation surrounding COVID-19. Together, the "unvax" and "spread" categories don't outweigh or overpower the hesitancy on vaccines and other pandemic mandates on Twitter, despite their large amount of positive sentiments.

Given these results and further analysis, the slight skew towards positive sentiment across the dataset is most likely due to the more or less even split between the negative and positive sentiments for each category. Topics, including "efficacy" and "mandate", have a pretty good distribution between the two polar attitudes. For example, the category "children" has the same percentage of 37% for both positive and negative sentiments, and the "political" category has a difference of about 7% between the positive and negative sentiments. Although "efficacy" and "mandate" categories, as mentioned before, also have this somewhat equal distribution of positive and negative sentiments, their categories have the most interaction with Twitter users; therefore, these categories and their slightly dominant negative attitudes have much more sway on the big picture COVID-19 conversation.

Regardless, the fact that the Twitter usership interacts with categories that have more negative attitudes charges and promotes vaccine hesitancy, not to mention a hesitancy to participate in any measures that are meant to curb — and hopefully eventually end — the COVID-19 pandemic.

Group Member Contributions

Leopoldo: data collection, open coding of 200 tweets, manual annotation of more or less 300 acceptable tweets, wrote Data and Methods section of final report

Xinyi: helped with data collection, open coding of 200 tweets, manual annotation of more or less 300 acceptable tweets, wrote Introduction and Results section, created visual representation of data, report formatting

Svetlana: helped with data collection, open coding of 200 tweets, manual annotation of more or less 300 acceptable tweets, data analysis (TF-IDF scores and other statistics), wrote Discussion section, final edits for report

References

Health-Canada. 2021. Health Canada authorizes use of Comirnaty (the Pfizer-BioNTech COVID-19 vaccine) in children 5 to 11 years of age. https://www.canada.ca/en/health-canada/news/2021/11/health-canada-authorizes-use-of-comirnaty-the-pfizer-biontech-covid-19-vaccine-in-children-5-to-11-years-of-age.html. Accessed: 2021-12-11.