Twitter Discourse of COVID in Canada

Claire Guyatt, Jake Furniss-Yesk, Nadia Hawarri
McGill University, Department of Computer Science
claire.guyatt@mail.mcgill.ca, jacob.f.yesk@mail.mcgill.ca, nadia.hawarri@mail.mcgill.ca

Introduction

The aim of this data science project is to gain an understanding of the key discussions currently happening around COVID in Canadian social media. To investigate this matter, data was collected from Twitter in the form of 1000 recent English Tweets, which were then manually labelled as pertaining to 1 of 7 topics as well as given an engagement and a sentiment score. Data analysis revealed that Tweets whose topic was less common would receive the highest levels of engagement, while Tweets about conspiracy theories tended to be the most negative. Overall, the findings of this project suggest that the public discourse around COVID in Canadian social media tends to be negative, but that it is fervent, as shown by the plethora of data to sample from and the total engagement from the original Tweets.

Data

A sample of 1000 original Tweets (retweets and replies were excluded) was collected from Twitter's API. The Tweets were collected from 3-5 days prior, to give ample time for engagement by other users; statistics show that the half-life of a Tweet is, on average, 18 minutes (*When Is My Tweet's Prime of Life?,* n.d.). Tweets were filtered for English only and were selected by the following case-insensitive keywords: 'COVID', 'COVID-19', 'vaccine', 'vaccination', 'Pfizer', 'BioNTech', 'Moderna', 'AstraZeneca', 'Janssen', 'Johnson & Johnson', 'J&J', 'Johnson and Johnson'. These keywords were chosen based on relevancy and the four vaccines that have been approved in Canada.

Collected Tweets spanned a period of three days, where 330 came from day 1, 330 came from day 2, and 340 came from day 3. Within each day Tweets were collected in chunks of 10 by a randomly generated hour, minute, and second, so the sample of Tweets would be spread out over the 24-hour period (e.g., there was no bias towards Tweets tweeted at 3 am or noon). A time within a day could not be randomly chosen twice to ensure there were no duplicate Tweets in the sample. A check was done to ensure there were no duplicates.

Methods

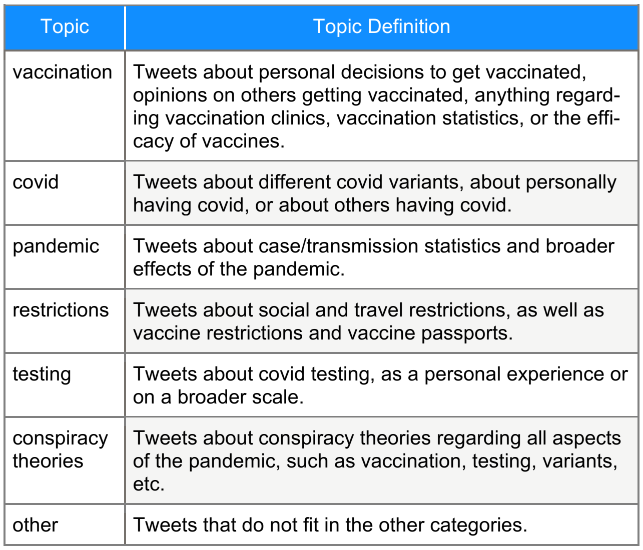
Data Collection

Once the sample of 1000 original Tweets was collected, the data was trimmed to include only the text, like count, quote count, reply count, and retweet count.

Data Annotation

A subset of 200 Tweets from the sample was used for open coding to determine central topics regarding the discourse around COVID-19 and vaccination in Canada. Three individuals open-coded the Tweets and came up with a list of topics; these lists were then cross-referenced to ensure topic validity, and matching topics were chosen for the final list.

From this open coding, the following 7 topics were generated: *vaccination, covid, pandemic, restrictions, testing, conspiracy theories*, and *other*. Tweets were labelled as conspiracy theories if they explicitly mentioned "conspiracy theory" or if they included "a belief that an event or situation is the result of a secret plan made by powerful people", as per the Cambridge Dictionary (Conspiracy Theory, n.d.), *and* the annotator was very confident the theory was not widespread nor had strong supporting evidence.



Topic	Topic Definition
vaccination	Tweets about personal decisions to get vaccinated, opinions on others getting vaccinated, anything regarding vaccination clinics, vaccination statistics, or the efficacy of vaccines.
covid	Tweets about different covid variants, about personally having covid, or about others having covid.
pandemic	Tweets about case/transmission statistics and broader effects of the pandemic.
restrictions	Tweets about social and travel restrictions, as well as vaccine restrictions and vaccine passports.
testing	Tweets about covid testing, as a personal experience or on a broader scale.
conspiracy theories	Tweets about conspiracy theories regarding all aspects of the pandemic, such as vaccination, testing, variants, etc.
other	Tweets that do not fit in the other categories.

Topic definitions from open coding of a randomly selected 200 Tweets from the original sample (n=1000).

All Tweets were then manually annotated to a single topic from the designed list, as well as were given a sentiment score of ‘positive’, ‘neutral’, or ‘negative’. Annotation was split evenly between 3 annotators to avoid fatigue. Finally, each Tweet was given an engagement score from the sum of its likes, retweets, quotes, and replies. Once topics and sentiment scores were added for each Tweet, the text was converted to all lowercase and punctuation was removed. Finally, common stop words (e.g. ‘the’, ‘and’, ‘I’ were removed from the data as well as numbers (excluding ‘19’)

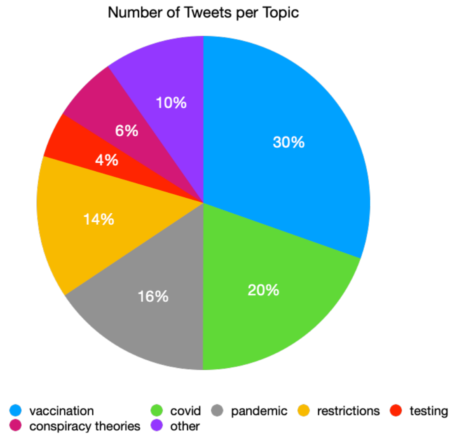
Data Analysis

Topics were characterized using term frequency – inverse document frequency (tf-idf). For each word occurring in any Tweet of the topic’s subset of Tweets, the frequency of the word (i.e., how many Tweets it appeared in) was divided by the total number of Tweets originally collected (n=1000). Using this method, the top 10 most used words specific to each topic were gathered. Topic sentiment was analyzed by subtracting the number of negative Tweets from the number of positive Tweets for each topic to give an overall sentiment score. Therefore, the lower the score, the more negative the sentiment around the given topic. The score was then divided by number of Tweets for the topic in order to make a fair comparison between topics. Finally, level of engagement was analyzed by summing the engagement score for each Tweet in the topic to give an overall engagement score, and once again dividing this number by total number of Tweets for the topic, resulting in the average engagement per Tweet for a topic.

Results

Topic Distribution

Manual annotation revealed that vaccination-focused Tweets were the most prevalent in the dataset (n=300), followed by covid (n=194), and then by pandemic (n=137).



Topic distributions by percent Tweets per topic.

Tf-idf

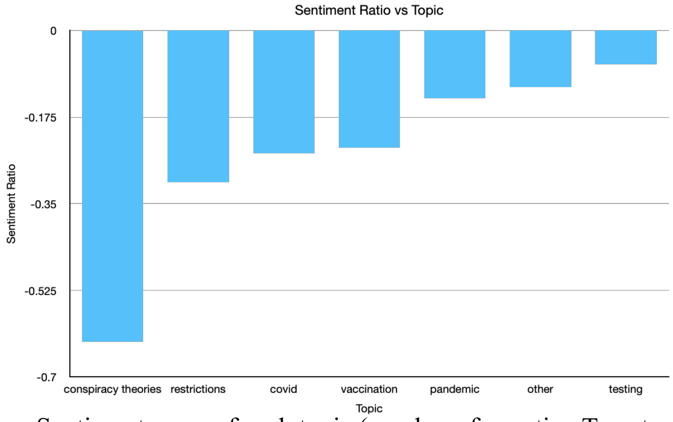
For each topic, the top 10 words calculated with tf-idf suggest what aspects the topic primarily concerns. The top 10 words for each topic correlate closely with their category, and the word covid appears as a stop 10 tf-idf word for each topic. A clear outlier is ‘lisa’ which has no obvious relation to the covid topic. An investigation into the outlier revealed that a celebrity lisa contracted the virus within the 3 days the data was collected. Similarly, in the vaccination topic ‘arm’ and ‘italian’ appear in the top 10 words, because on one day of data collection, an Italian man attempted to avoid the vaccine with use of a fake arm. We left these outliers in the top 10 list to demonstrate that though most words (and therefore common concerns for each topic) remain constant, the discourse is also continually affected by singular events in the world.

Topic	Top 10 Words
vaccination	vaccine, moderna, dose, covid, pfizer, arm, covid-19, vaccination, doses, italian
covid	covid, omicron, variant, covid-19, delta, lisa, cough, guy, infections, detects
pandemic	covid, covid-19, rate, sales, neutral, updates, omicron, deaths, cruise, active
restrictions	covid, vaccine, restaurants, war, plan, enforce, stores, ban, passports, restrictions
testing	tests, test, covid, testing, campus, raploch, lateral, flow, kits, arrivals
conspiracy theories	covid, channels, bla, liability, tyranny, yearly, massive, scam, century, crooked
other	covid, rs, lakh, modi, covid-19, team, orbison, royals, deer, challenge

Top 10 words for each topic calculates with tf-dif (number of Tweets the word appears in for a given topic / number of times the word appears in the total sample of Tweets (n=1000)).

Sentiment

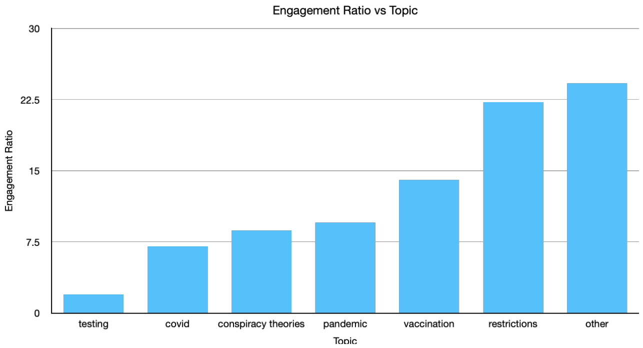
Sentiment scores show that each topic was more likely to be labelled as negative than as positive. Conspiracy theories-focused Tweets were the most likely to be labelled as negative over positive, more than double as likely as the following topic, covid. Of the 7 topics, testing-focused Tweets were the least likely to get a negative label over a positive label.



Sentiment score of each topic (number of negative Tweets - number of positive Tweets) relative to the number of Tweets in the topic category.

Engagement

Engagement scores show that Tweets in the ‘other’ category received the most engagement from other users, followed closely by restrictions-focused Tweets, while testing-focused Tweets received the lowest levels of engagement from other users.



Engagement score of each topic (by likes, retweets, quotes, and replies) relative to the number of Tweets in the topic category, i.e., average level of engagement per Tweet for a topic

Discussion

Unsurprisingly, the three most general topics (vaccination, covid, and pandemic) made up the bulk of the sample. However, while they represented the most Tweets by volume, their per Tweet engagement and sentiment were unexceptional, ranking at most third in both. This suggests that the discourse around the general or commonly brought up aspects of the pandemic (including the status of the pandemic, the virus itself, and existing vaccines) while still the largest topics, do not have the most activity or controversiality on the platform. This could be due to the settling of people’s opinions, and their losing interest in repeated arguments.

Through annotation of the data, it became apparent that Tweets pertaining to restrictions or other niche topics would receive more engagement per Tweet. Surprisingly, the data revealed that Twitter users were 59% and 79% more likely to engage with Tweets about restrictions or other topics compared to vaccination. This may be because Tweets about restrictions generally came in the form of surprising or salient news updates, such as news about contentious restrictions on the Australian COVID quarantine compounds whereby travellers are supposed to isolate for 14 days. Furthermore, Tweets that fall under the ‘other’ topic were more event-based and by definition of their topic, unique, meaning that they were the most relevant to the moment and therefore also receive more engagement.

It is important to note, when interpreting results regarding sentiment, that sentiment scores do not reflect general sentiment for a given topic. Rather, the results reflect the sentiment surrounding the discourse of a given topic. For instance, though the vaccination topic has a sentiment score of -2.86, that does not imply that people view vaccination negatively. What it does imply is that the discourse surrounding vaccination is more likely to be negative than positive, whether that be, for example, negativity towards people getting vaccinated, or negativity towards people not getting vaccinated.

Conspiracy theory focused Tweets had the most negative sentiment score per Tweet by far with a margin of over 105% compared to restrictions, the second most negative topic, suggesting that sentiment of discourse surrounding conspiracy theories is the most negative. This seems a natural outcome of the inherent negative quality of conspiracy theories per the topic’s definition, oriented towards distrust in organizations and leaders, as well as the promotion of misinformation. These Tweets would then inherently trend towards a negative outlook of current events, in a pessimistic dissent of popular belief, or in pessimism towards those mistrusting beliefs. Interestingly, these Tweets only represent 6% of our sampled data, and would only receive a moderate amount of engagement, ranking third to last with a score of 8.677 (61% less than the afore-mentioned restrictions). This suggests that while people Tweeting conspiracy theories are most likely to express a negative sentiment, they don’t necessarily receive widespread support, and represent a highly negative, dispersed, and fractured minority of Twitter users.

Finally, testing proved to be an all-around un-engaging topic for Twitter users. It represented the smallest percentage of our sampled Tweets, scored last in per Tweet sentiment and engagement, and had the most scattered words in its ten highest scoring tf-idfs, including orbison and challenge. This may be explained by its diminishing relevance throughout the pandemic. Whereas early in the pandemic, testing may have been a highly relevant topic as the sole metric and method of combating the spread of COVID-19, at this point vaccination is clearly a more relevant metric or point of contest for Twitter users. Simply put, at this point in the close to 2-year long pandemic, people are both used to testing, and have greater issue with vaccination or restrictions.

In closing, we can note limitations in our data collection and the design of our topics that lent to some unavoidable biases in the annotation of our sample data. The main limitation in our data is that it may contain data from elsewhere in the world than Canada. Our only measures to focus on Tweets from Canadians or Tweets within Canada were our keywords, where we focused on Canadian vaccines, as well as limiting to only English Tweets. These measures, however, would not reliably filter out Tweets from other English-speaking parts of the world.

Regarding topic validity, we found it sometimes difficult to separate conspiracy theories from the other topics. By our definition, a Tweet would be considered pertinent to a conspiracy theory if it explicitly mentioned "conspiracy theory", or more likely if it included "a belief that an event or situation is the result of a secret plan made by powerful people" and the annotator was confident the theory was not widespread nor had strong evidence in support. However, given the current politicization of fact, subjective realities have become all too prevalent. This directly translates to difficulty in objective categorization of conspiracy theories, as it can become difficult to distinguish an argument that is using a less reputable source of information versus a source of information that is disconnected from reality. As annotators, we also must account for our own bias in judging sources of information, and account for our own backgrounds. This naturally lends to inherent bias in judging the topic of conspiracy theories. To combat this possible limitation, we frequently checked Tweets of this nature with other annotators and attempted to go off precedence in subsequent cases.

Secondly, due to the complexity of the matter, categorizing Tweets about a global pandemic and all it entails can be difficult to separate cleanly into distinct categories. Tweets can easily have complex crossover between topics. For example, it could be argued that vaccine mandates belong in the restrictions topic, as it's governmental policy that informs what people are able to do. However, people frequently contest vaccine mandates in conjunction with 'evidence' or arguments against vaccine efficacy, and therefore could also be attributed to the vaccination pool of Tweets. But if the evidence is problematic and untrustworthy, or there's also a tone of distrust towards the government, the Tweet could be interpreted as belief in a conspiracy. With a plethora of examples just like this one, it proved inherently difficult to distill complex opinions and pieces of information into segregated topics, relying on the best judgement of the annotators, and therefore creating inherent bias in the data.

Member Contributions

Claire

Came up with annotation topics, annotated the last 340 Tweets, then cleaned up the annotated data for analysis and did the original data analysis script for engagement and sentiment scoring. Wrote the intro, data, and methods sections of the report, and added to/edited the results and discussion sections. Put all the data, figures, and written sections together into the final report.

Nadia

Helped with tweet collection using the twitter API and conversion of JSON data to CSV. Helped come up with annotation topics and annotated the middle 330 tweets. Wrote the results section of the report and created the graphs and tables to highlight the results.

Jake

Contributed through the creation of the Tweet collection script, including app API authentication, utilizing Twitter's API query utility, and then performed the sampling. Then proceeded to convert JSON data to a CSV, configured to a Google Sheet with data validation for consistent annotation, and annotated the first 330 Tweets. Finally, built off the data analysis script, adding per tweet averages and tf-idf calculation/ranking, and then wrote discussion about the findings.

References

When Is My Tweet's Prime of Life? (A brief statistical interlude.). (n.d.). Moz. Retrieved December 9, 2021, from https://moz.com/blog/when-is-my-tweets-prime-of-life

Conspiracy theory. (n.d.). Retrieved December 10, 2021, from https://dictionary.cambridge.org/dictionary/english/conspiracy-theory