COVID-19 in Canada, Late-October 2021

Mohanna Shahrad, Maggie Shao, Yann Bonzom
McGill University

Introduction

With the COVID-19 pandemic now completing its second full year, it is truly incredible how much it has affected seemingly every aspect of society. From large-scale effects such as increased unemployment to more personal effects such as reduced socialization, this pandemic has made a significant dent. In this research project, we explore COVID in the context of late-October Canada with the aim of seeing what the salient topics are along with what the general population thinks of them.

We began our exploration by extracting over 10 000 COVID-related tweets from Twitter, a platform that is especially popular with people who seek to voice their (often rather strong) opinions. We then categorized over 1000 tweets based on what they most relate to, with our final set of categories being Policy, Pandemic Effects, Vaccination, Infections, and Science. Additionally, we ascribed either a negative, neutral, or positive sentiment to each tweet so we can explore the Canadian population's opinions regarding these categories. With this now-annotated data, we performed both frequency and tf-idf score analyses to determine the most significant words appearing in the tweets for each category, and we also calculated negative/positive sentiment ratios to assess public perceptions.

From our results, we were able to uncover some of the most salient topics surrounding COVID in Canada in late October. Most apparently, the recent approval of COVID vaccinations for children aged 5 to 11 years old has caused quite a stir on Twitter, though from what has been seen during the annotation step, this has left many parents very much relieved. Nonetheless, as was made clear from the Science category's analysis, there remains some (though statistically not too significant) vaccine hesitancy as well as distrust in the studies published by vaccine companies. Furthermore, there is significant discussion surrounding the vaccination rollout (in particular with regards to booking appointments (including for children) and 3rd booster shots) as well as policies such as vaccine passports, testing requirements, and travel regulations. General discontent regarding the whole situation is, however, clearly apparent – but this is only to be expected given that this is a pandemic.

Data

To collect our tweets, we used Tweepy and the Twitter API and filtered them by both words and hashtags containing at least a word from the following: covid, corona, vaccination, Pfizer, Moderna, AstraZeneca and Janssen. Then, to make sure the tweets were mostly from Canada, we selected them from within a 1500km radius of two arbitrary geolocations near the center of the country. The two regions cover the entirety of the country, as well as a small area of the United States from where the tweets will be removed during the data cleaning process. Hence, we collected over 1000 tweets for each of the filter words and each of the two regions, leaving us with 10 000 raw tweets to work with. We then formatted our data as a CSV file with the columns: 'tweet', 'created at' and 'location'. These are useful information to ensure, during the data cleaning process, that we only keep the tweets posted within a three-day window and in Canada.

With 10 000 COVID-related tweets, it was time to clean up our raw data. First, we removed any duplicates based on the text content of each tweet. Since we had a substantial number of posts, we could afford to drop all retweets given that they mostly cover the same topics already discussed in the original tweet. Then, to ensure that the tweets are strictly Canadian, we filtered out any posts that, for their location, did not contain any words that appeared in a list of Canadian provinces and their abbreviations as well as Canada and CA. Next, we applied a simple filter to keep only tweets posted within a 3-day period. However, not all posts remaining were relevant for the analysis. This might be caused by several factors such as posts containing COVID-related terms yet actually being unrelated to our topics, as well as posts that are incomprehensible without details of the conversation's context. Consequently, since removing such tweets would have to be done manually during the annotation process, we kept 3000 tweets so we can skip any posts that are unclear or obviously unrelated as we annotate 1000 of them. Then, we formatted our cleaned dataset as a CSV file containing only the 'text' column.

With this list of tweets, we developed our typology and started to annotate them, exchanging questions on odd cases and thereby further refining our definitions for each category. During our labeling process, we found 191 posts that

were either unrelated to our topics or unclear. We performed this annotation using Google Sheets so we could all work on the same dataset and used simple Sheets tools to remove all unnecessary rows and create the final annotated dataset suitable for analysis. In the end, we kept 1025 correctly labeled tweets in our final dataset which we formatted as a CSV file with 'category', 'sentiment' and 'tweet' columns.

Methods

To optimize the computation of the tf-idf scores, we tweaked our final dataset by the following preprocessing steps. Since we want to avoid having redundant words with different cases, we treated each word as case insensitive by converting all the text to lowercase. Then, we expanded all occurrences of most English contractions (e.g., changing 'don't' to 'do not') to further standardize our text. Furthermore, as the tagged usernames that appear in tweets do not provide significant insights for the analysis, we decided to remove them (done by removing all words starting with '@'). Similarly, we decided to not consider hashtags (by removing all words starting with '#') and thereby exclusively focus on the raw text components of the tweets. The last and most crucial step was the removal of stop-words, which are the most frequently used words in a language and consequently usually do not carry important meaning (e.g., words such as 'to', 'from', 'the', etc.). To achieve this, we used the Natural Language Toolkit (NLTK) and its list of English stop-words. We also experimented with stemming and lemmatization techniques with the goal of reducing inflectional forms and sometimes derivationally related forms of a word to a common base form using NLTK (Beri, A., 2021). However, these methods resulted in less meaningful words at our final analysis and this step was therefore excluded from the text preprocessing phase. Finally, we filtered out punctuations, digits, and numbers (since those terms are not considered words) and to keep our text data well-formatted, we removed any extra spaces left between words by the previous preprocessing steps.

Next, we moved forward with computing the tf-idf score of the words in each category as a way of evaluating the words relevancy to each category. We decided to not only look at individual words, but at pairs of words as well. This decision was made mainly because pairs of words might tell us more about the content of the tweets compared to single words, which turned out to be true. For this reason, we used n-grams, i.e., neighboring sequences of items in a document of length n (Dios, E. C. D. 2020). For example, the 2-grams of 'COVID is bad' are '(Covid, is)' and '(is, bad)'. We used an n-gram range of (1,2) in our analysis (meaning we look at single and pairs of words) and then computed the tf-idf score of the words and pairs together in addition to using only single words. This resulted in our lists of highest tf-idf scored words that can be found in Table 1 and Table 2.

Lastly, we performed a sentiment analysis on each category. In addition to visualizing the percentage of sentiments by category (Figure 1), we also compared the categories by defining the ratio of negative sentiments over positive ones to be able to compare general opinions on different categories. The results of this analysis can be found in Figure 2.

Results

Developing Typology

Producing an effective typology, or list of mutually exclusive categories, proved to be quite difficult. However, following a first two-hour call in which we did an open coding of over 200 tweets, followed by some individual data annotation, and yet another call, we were able to decide and define our final typology:

- *Policy*: related to any policies put in place or discussed by a body (government, a company, etc.). Here, such policies often concern themselves with vaccine passports, vaccine approvals, workplace restrictions, and travel restrictions.
- *Pandemic Effects*: related to the pandemic itself and its consequences including, for instance, social, economic, educational, and health-related effects.
- *Vaccination*: related to the COVID vaccine rollout and the act of getting vaccinated, excluding vaccine-related policy and the scientific aspect of vaccines.
- *Infections*: related to any information on the tracking and spread of COVID, including statistics on new cases and deaths as well as information regarding testing.
- *Science*: related to scientific findings, research, studies, and explanations regarding COVID and vaccination.

Statistics on Annotations

Figure 1: Sentiments by Category

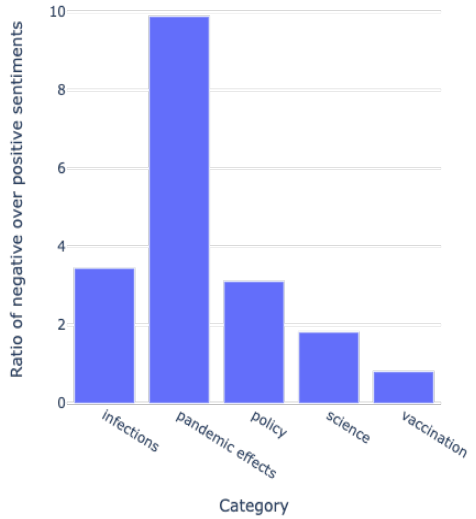

Figure 2: Ratio of Negative Over Positive Sentiments

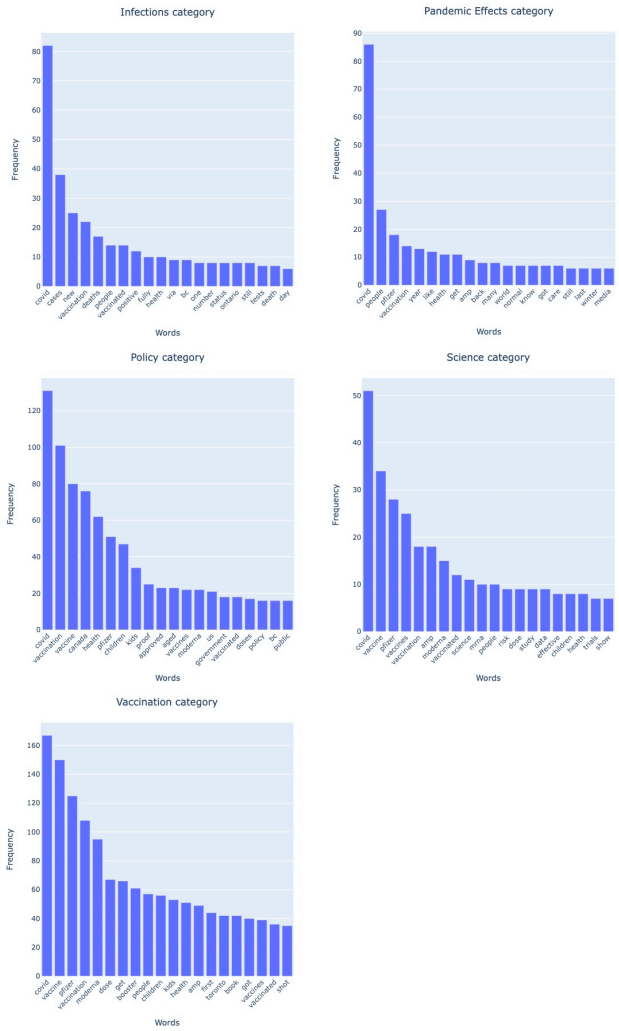

Figure 3: Highest-frequency Words by Category

Policy	Pandemic Effects	Vaccination	Science	Infections
proof	stocks	rd	study	(new, cases)
(vaccine, children)	operation	spots	cell	(cases, covid)
(health, Canada)	(many, people)	detailed	type	(covid, deaths)
aged	dark	availability	antibody	(mp, tests)
(proof, vaccination)	airborne	(Toronto, health)	mrna	(tests, positive)
approves	shipping	(spots, added)	rna	(positive, covid)
Authorized	violence	(added, first)	contradictions	erin
(mandatory, vaccination)	guilty	(first, available)	adolescents	(fully, vaxed)
(children, aged)	piece	(available, date)	mixing	(erin, otoole)
(Canada, approves)	women	(date, book)	(ppl, dose)	(reporting, new)

Table 1: Words with Highest tf-idf Scores (n-grams (1, 2))

Policy	Pandemic Effects	Vaccination	Science	Infections
proof	stocks	rd	study	erin
aged	operation	spots	cell	positive
approves	dark	detailed	type	mp
authorized	airborne	availability	antibody	tory
asks	shipping	aged	mrna	spotlight
government	violence	appointments	rna	loss
provincial	guilty	book	contradictions	blah
approve	piece	Toronto	adolescents	dogs
policy	women	date	mixing	eve
authorize	economic	arm	immune	reporting

Table 2: Words with Highest tf-idf Scores (Single Words)

To better understand the nature of our final clean data, we created visualizations showing the top twenty most frequent words in each category. The results of this initial data analysis can be found in Figure 3. As expected, words such as "covid", "vaccine", and "vaccination" were the most frequent ones in all categories. However, there were also several unique words in each category. After having initial observations on distinct categories of data, we performed the tf-idf analysis, with the results presented in Tables 1 and 2.

Discussion

Now that we have the results extracted from our annotated data, we can proceed by interpreting them topic by topic.

Policy

From Table 1, it is immediately evident that the most salient topic regarding policy in Canada surrounding COVID during our data-collection period was the government's approval of vaccinations for children aged 5 to 11 years old. This can be seen from the word pairs '(vaccine, children)', '(children, aged)' (which usually occurs in the context of specifying the age range of children that can now get vaccinated), and words regarding this approval such as 'approves', 'authorized', and '(Canada, approves)'. Additionally, the most frequent words in this category include 'children' and 'kids', further backing this.

Previously, only people aged 12 and up were allowed to get vaccinated against COVID – with this new law, an even larger portion of Canada's population can now get vaccinated. Consequently, it is only to be expected that a lot of online discourse will surround this new policy.

Another important topic in this category is vaccination proofs, as is evident from Table 1's elements of 'proof', '(proof, vaccination)' and '(mandatory, vaccination)' as well as the frequency table's 'proof' column. Vaccination proofs, often in the form of vaccine passports, are required in most provinces to partake in any discretionary pastimes such as going to restaurants and theaters. Because of their sheer impact on our freedom, it is to be expected that this is currently a major topic.

Policy, per Figure 2, has the third-highest negative/positive sentiment ratio at around 3.1. This indicates general discontent with government and company policies, though it is difficult to infer much from this value given the variety of topics in this category. From our annotation process, it was evident that there was general skepticism towards vaccinating children, especially with regards to health concerns.

Pandemic Effects

From Tables 1 and 2, it is perceptible that most topics in this category expressed negative concepts such as 'violence' and 'guilt'. The terms in this category's results represent a diverse range of the pandemic's consequences. For instance, the term 'stocks' refers to the economic impacts of the pandemic and more specifically relates to two ongoing topics: the pandemic's winners in the stock market (such as Pfizer and other pharmaceutical companies) and the major stock market crash such as with oil stocks.

Additionally, the term 'violence' refers to the negative social impacts of the pandemic. It is about the growing concern about the social crisis created by the pandemic such as COVID-related violence among different communities. As another example, the term 'shipping' can be considered as a negative consequence referring to people experiencing shipping delays caused by COVID-19.

The Pandemic Effects category, based on Figure 2, has the highest negative/positive sentiments ratio, which is a strong indicator of Canadian general dissatisfaction with the consequences of the pandemic including social, economic, and academic effects. (The negative/positive ratio of this category is about three times greater than the Infections and Policy categories). Nearly 63% of the tweets in this category carried a negative sentiment regarding the difficulties people have been experiencing due to the pandemic.

Vaccination

As is apparent from the results of Tables 1 and 2, the most significant topic in this category is about vaccination appointments (including booking the appointments or the available spots/newly added spots at vaccination centers). This can be concluded from the existence of terms such as 'spots', 'book', 'available', and 'date', and was also often seen during the annotation process. The tweets with such content are usually either individual personal experiences in taking their vaccines or advertisements posted by the vaccination centers.

Another term visible in the results of the Vaccination category is 'rd'. These terms are the remaining parts of the terms '3rd' in the dataset after applying the preprocessing

step of removing numbers, which is clearly a minor limitation of this approach. The significant presence of this term relates to the ongoing topic regarding the administration of 3rd booster shots of COVID vaccines. This is done with the aim of fighting against new COVID variants such as Omicron. More specifically, the tweets were either personal experiences on getting third doses or to encourage people to get their third shots as soon as they are eligible.

Vaccination, based on Figure 2, has the lowest negative/positive sentiment ratio. Most of the tweets in this category (49%) had a neutral sentiment, likely due to the many advertisements for vaccination clinics. Among the remaining tweets, the proportion of positive and negative sentiments are quite similar (22.7% & 28.3% respectively), which could indicate that people's perspectives regarding the vaccination process are segregated.

Infections

Most posts under the infections category share general statistics on positive COVID tests and its related deaths. This can be seen in Table 1 where the groups of words with the highest td-idf scores are '(new, cases)', '(cases, covid)', '(covid, deaths)', '(tests, positive)'. In addition, a considerable number of posts discuss specific newly reported COVID cases, notably in institutions or within a government body, as we can see by the terms '(mp, tests)', 'erin', '(erin, otoole)' in Table 1, as well as the words 'erin', 'mp' and 'tory' in Table 2. It is important to note that here, 'mp' is referring to members of Parliament. Furthermore, any tweets with mentions of Erin O'Toole, the leader of the Conservative party, or of Tory, the nickname of this party, talk about various members of Parliament having recently tested positive for covid. We can see that all the previously presented discussions around this topic have a neutral sentiment since they are merely conveying information and facts on new Covid cases. We can also see this in the Figure 1 which shows that 61.5% of posts have a neutral sentiment.

The remaining posts have a negative to positive sentiment ratio of 3.4, as shown in Figure 2. This large disparity might be because the number of COVID cases does not seem to be improving and the number of deaths has been increasing. However, some try to stay positive by promoting vaccination as a solution to reduce the spread of covid. We know this by looking at Figure 3 where 'vaccination' is shown to be one of the most frequently used words in this category.

Science

The Science category, as expected, contains a lot of terms related to the vaccines, their properties, and their safeties. This is evident from Tables 1 and 2, which contain terms such as 'cell', 'antibody', 'mRNA', and 'immune'. Additionally, there is clearly a significant amount of discussion related to the results of recent studies, as indicated by the term 'study' having the highest tf-idf score in both Tables 1 and 2. The presence of the term 'adolescents' clearly relates to the newly implemented policy of vaccinating the younger population, with such posts often being in relation to the safety concerns of this new decision. Finally, there is also discussion related to 'mixing' COVID vaccines, which was, during the annotation process, often found to be in relation to booster shots that Canada has recently made available.

Science, which given its objectivity would be expected to have a negative/positive sentiment ratio of around 1, surprisingly still has a relatively high score of 1.8. This might be a consequence of people's mistrust in the results of COVID-related studies as well as people hesitating to vaccinate their children and get booster shots themselves. Lastly, during annotation, we have found several occasions of scientific fake news which were categorized as 'Science' and often carried a negative sentiment– this might help explain this high negative/positive sentiment ratio.

Concluding Notes and Limitations

From the above analysis, there clearly remains more to find out about these salient topics. Given the nature of the sentiment analyses, it is difficult to pin down exactly what causes (dis)contentment in each category, since each category contains several topics. Going further into this and exploring the sentiments with regards to each of these subtopics could help reveal additional insights.

As for the approach of using n-grams in our tf-idf calculations and thereby generating Table 1 in addition to the single-word Table 2, this has proven particularly fruitful. Especially in certain categories, such as Policy, using only single words prevents us from seeing a lot of information. The word pairs, such as for instance '(vaccine, children)', are much more valuable to our analysis.

Finally, the frequency tables has proven very useful since some words, despite being common across most categories, do carry meaning and ended up being useful to our analysis.

Group Member Contributions

Maggie Shao worked on collecting the data, including figuring out how to collect mostly Canadian tweets, ensuring an adequate number of COVID-related tweets, and organizing them for preprocessing.

Yann Bonzom performed the data preprocessing, organized the Google Sheet used for data annotation, and processed that Sheet to create the final annotated data file.

Mohanna Shahrad calculated the tf-idf scores, created graphs to visualize our data, and performed an n-gram approach to analyzing the tweets to extract further insights.

We all helped one another at every step, called regularly (especially while developing our typology and discussing our results), and performed the data annotation and report writing in equal parts.

References

Beri, A. (2021, January 27). Stemming vs lemmatization. Medium. Retrieved December 12, 2021, from https://towardsdatascience.com/stemming-vs-lemmatization-2dad-dabcb221

Dios, E. C. D. (2020, May 31). From dataframe to N-grams. Medium. Retrieved December 12, 2021, from https://towardsdatascience.com/from-dataframe-to-n-grams-e34e29df3460.