Quantitative analysis of Twitter discourse about COVID in Canada

Alon Shapiro[1], **Hugo Baraer**[1], **Luca Weishaupt**[1]

[1]McGill University
845 Sherbrooke Street W Montréal, Quebec H3A 0G4
alon.shapiro@mail.mcgill.ca, hugo.baraer@mail.mcgill.ca, luca.weishaupt@mail.mcgill.ca

Introduction

Over the past 21 months, the COVID-19 pandemic and the imposed restrictions have had significant impacts on Canadians' daily lives (Haleem, Javaid, and Vaishya 2020). New waves of infection are occurring as countries roll out vaccines, and new mutations of the disease are being discovered. Canada is said to be a leader in immunization and mitigation of the spread of COVID. However, the pandemic has also become a polarizing and political issue (Crabu et al. 2021), which is reflected in the discourse that can be observed on popular social media platforms such as Twitter.

A study conducted in Spain (Santoveña-Casal, Gil-Quintana, and Ramos 2021) shown that studying the citizen's feelings about COVID helps the interaction between legislators and citizens, hence improving adherence to imposed sanitary measures. Therefore, organizations and legislators must know how citizens in Canada feel about COVID. The purpose of this study was to characterize and classify the discussions about COVID in Canadian social media and to study the general response of Canada's population to the pandemic and vaccination. To do so, 1100 tweets concerning COVID were collected over a three day window, from November 24th 2021 11:59 PM, to November 27th 2021 11:59 PM.

The collected tweets were open coded to establish some relevant categories, which all the tweets were then categorized into. Tweet sentiment was also labelled as positive, negative, or neutral. Each category was then analyzed to identify the salient discussion topics with TF-IDF analysis. The most common discussion present on Canadian social media was found to be a critique of response to COVID by the government and vaccine safety and efficacy, which was mostly negative. The second most common discussion surrounded direct news articles posted by known news networks and government bodies and overwhelmingly carried neutral sentiment and were meant to inform the public.

Data

A Twitter scraper was implemented using a python module called tweepy. 1100 tweets were scraped over a range of three days: from November 24th, 2021, 11:59 PM, to November 27th, 2021, 11:59 PM. Using the twitter API, tweets from each day were randomly scraped from random Canadian Twitter users to avoid any bias. Furthermore, to ensure that the data set was a representative sample of the discourse about COVID in Canada, the tweets that were scraped had to meet the following criteria:

1. The tweet is written in English.
2. The case insensitive version of the tweet contains the word ”covid” or ”coronavirus”, or the name of one of the four approved COVID vaccines: ”Pfizer-BioNTech”, ”Johnson&Johnson”, ”AstraZeneca”, and ”Moderna”. Abbreviations ”Pfizer” and ”J&J” were accepted.
3. The tweet is unique and is not present in the data set. This includes retweets.
4. The tweet has been posted by a user with a location tag within Canada.

Condition 4 was checked by passing the user's location tags into a Google Maps API to determine what country the location falls in. We acknowledge that the limitation imposed by condition (4) may bias the data set to only include users who are well acquainted enough with Twitter and don't mind publicly sharing their location. After cleaning the data as outlined in the methods section, 1044 tweets remained in the final data set used for analysis.

Methods

Open Coding

Every team member started with 200 unique tweets from the data set and conducted open coding to develop a comprehensive list of categories to classify the tweets. Next, all members met and merged their initial categories into broad categories that revolved around the intent or nature of the tweets. Finally, 200 new tweets were labelled with the merged categories to check their validity and establish firm rules for choosing which category.

A category for items that should be removed was used for manual data cleaning during annotation. E.g. a tweet that only contains the word ”COVID” might be labelled for removal since it carries no information about sentiment or about which category it best fits into.

Category Annotation

A hand-crafted annotation tool was used to display the original tweet in an embedded web page to provide more contextual information, such as showing possible interaction with other tweets or the attached images. Using this tool, tweets were labelled with a sentiment and the category they fit into best. All three team members annotated the original set of tweets independently, resulting in three labels per tweet.

Sentiment Annotation

Generally, tweet sentiment was annotated as positive if it expressed themes of hope, support, tolerance, admiration, optimism, enthusiasm, appreciation, compassion, love, admiration, or any pleasant/desirable emotions. Examples include tweets with a hopeful outlook regarding the pandemic, support and tolerance towards sanitary measures, admiration of healthcare workers, and working from home enjoyment.

Tweets that expressed themes of criticism, harm, accusation, manipulation, pessimism, anger, sadness, condescendence, incomprehension, overwhelm, tiredness, hate, deception, or tweets expressing unpleasant/undesirable emotions were annotated as negative. Examples include posts against sanitary measures, posts suggesting some harmful conspiracy, pessimist outcome to the pandemic, or sadness in touching a personal story.

Tweets that did not display any of the above characteristics, lacked in emotions or showed a mix of positive and negative emotions were labelled as neutral. Some examples include: statistics, mitigate opinions, neutrally reported news, and emotionless opinions.

Merging Labels

All labels were merged using majority voting, implemented in a python script. I.e., when two or more observers agreed on sentiment or category, that classification was kept. In case of a disagreement in sentiment between the three team members (1 Negative, 1 Neutral, 1 Positive), the tweet was labelled as Neutral. For the categories, the tweets showing a disagreement between the three observers on its category were set aside. The category for these tweets was then discussed and discussed with all three members.

Tokenization and TF-IDF

To analyze discussions topics, words presented in the tweets were analyzed. We tokenized our text corpus by first finding individual words separated by white spaces or non-alphanumeric characters. Next, all stop-words were removed, like stop words such as ”and” or ”I” carry little significance for discerning the linguistic intent of a tweet. Stop-words were taken from a list of NLTK's English stop words (complete list available here). The stop word provided in homework eight was not used; a certain word with a connotation (such as get) was included in the list and judged helpful in the comprehension by the team. Next, words were stemmed using a Porter stemmer so that terms that conjugated verbs and singular or plural nouns are not treated as separate words. For example, the words vaccine and vaccines were treated as the same word. Word associations made by the porter stemmer are presented in the Results section.

Finally, the term frequency times inverse frequency (TF-IDF) was computed by treating each sentiment and each category as a document. Additionally, statistics about the number of tweets in each category were computed. Finally, to avoid biased TF-IDF calculation, words used to select the tweets during the data-gathering steps such as 'COVID' and the various COVID vaccine brand names were excluded.

Results

Chosen Categories

After merging the labels, five categories remained that are shown in Table 1. These categories were designed to be as objective as possible. Informative topic includes tweets posted by news organizations which contain published news articles, informational tweets posted by a verified account of an organization such as a department of the government, police force of a certain area within Canada, etc. It also contains tweets that state information that can be verified relatively easy, such as the latest infection numbers.

Topic	**Description**
Informative	Presenting easily verifiable information
Questions	Genuinely asking for information
Sarcasm	Humorous or disingenuous remarks
Critique	Opinions and critical remarks
Personal	Testimonies and sentiments

Table 1: Categories chosen after merging

Information which would be harder to retrieve by the general public from a reliable source, or that requires domain knowledge rather than general knowledge such as understanding the science found in published research, or claims that can fall under a conspiracy belief were labelled as Critique, as those were considered personal opinions. The questions topic only included genuine questions (e.g. 'does anyone know the current restrictions in Montreal?') looking for objective answers. Rhetorical or sarcastic questions and questions which don't have an objective answer (e.g. 'how long can you wear a mask before it starts annoying you?') were not included in the category. The sarcasm topic included any tweets which were written as a joke, or had a humorous intent. The critique topic included any tweets criticizing actions taken by politicians, government bodies, police, scientists and healthcare providers directly related to COVID, as well as opinions about the pandemic, sanitary measures, and COVID vaccinations. Any tweets expressing belief in conspiracies related to COVID were also included in this category. Lastly, the personal topic included any tweets that expressed a certain emotion, or a personal testimony or account of an event or story that happened to the author of the tweet. For example, sharing about a family member's COVID infection or expressing mental fatigue from worrying about the pandemic would fall under this category.

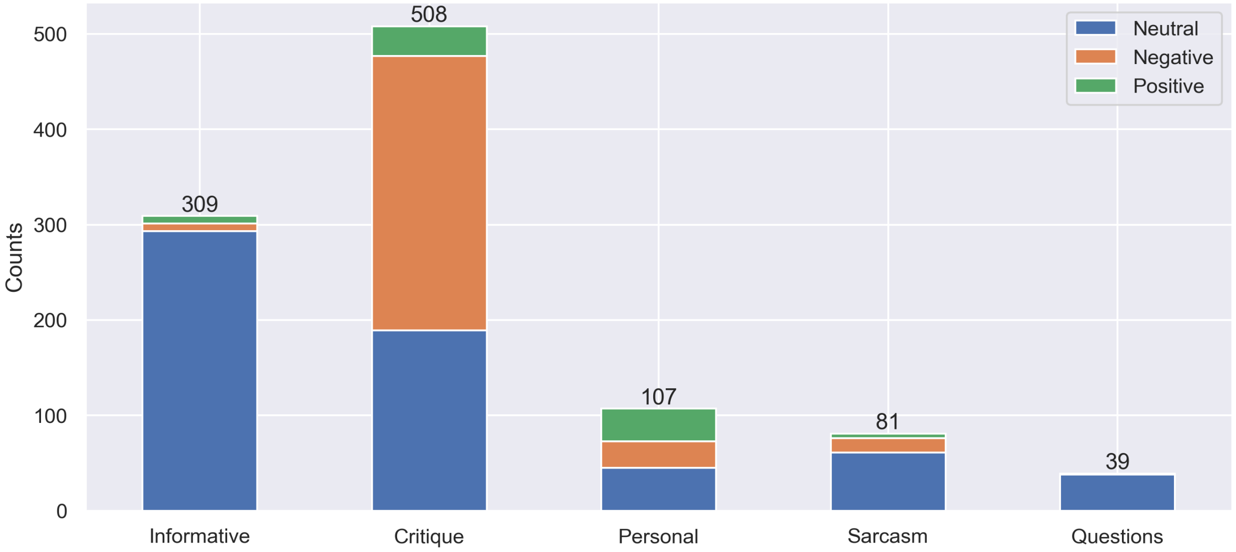

Figure 1: Sentiments make up by tweet category. Each of the six categories is represented by a bar, with the total number of tweets in the category presented above the bar. The proportions of sentiments in each bar are indicated by their colors.

After filtering out posts marked for removal during data annotation, 1044 tweets remained in the final data set. From the initial 1100 tweets, only 31 tweets (<3%) could not be merged using majority voting, indicating that the defined categories minimized subjectivity. For the sentiments, only 20 tweets were in complete disagreement. The low inter-observer variability demonstrates low subjectivity in the sentiment definition and categories, which establishes confidence in the annotation process and chosen typology.

Category and Sentiment

Figure 1 presents the distribution of tweets' category and sentiment. Firstly, analysis shows that the most abundant category of COVID tweets present is critique and opinions, with 48.6% (508/1044) of sampled tweets. This shows that Twitter is a significant source of citizen's opinion. Most of the tweets in that category is having a negative sentiment, with 57% being of negative sentiment, 37% neutral and 7% positive. This shows that the critique towards the current situation is mostly negative. Secondly, the second most abundant COVID discussion topic is informative tweets which represents 30% of the sampled tweets (309/1044). The vast majority, 95%, are of neutral sentiment, while negative sentiment and positive sentiment represent 3% and 2% respectively. Thirdly, 10% of the sample (107/1044) of the tweets are in the personal category. Sentiment distribution is the most uniform in this category (42% neutral, 25% negative, and 33% positive). Fourthly, the sarcasm category represent 8% of the distribution (81/1044), with a percentages of: 75% neutral, 18%negative, and 6% positive.

Looking at sentiment distribution over all categories : 7.5% (78 tweets) have a positive sentiment, 32.5% (340 tweets) a negative one, and 60% (626 tweets) a neutral one. Neutral tweets composed the majority of the sample, with a positive:negative:neutral ratio of 1:4:8.

TF-IDF

In the general case, the porter stemmer considered words being of the same meaning only for the plural form of nouns (eg. peopl stands for people and peoples). Some exceptions occurred and are presented in Table 2.

Root	words originating from root
vaccin	vaccinated, vaccinate, vaccinations, vaccination, vaccines, vaccine, vaccinating
get	getting, get, gets
report	reports, reporter, report, reporting, reported
test	tested, testing, tests, test
need	need, needs, needed, needing
make	make making makes
name	names, name, naming, named
stop	stop, stops, stopped
mask	mask, masking, masks

Table 2: Expanded words that were stemmed by the Porter stemmer

Figure 2 presents the finding for the TF-IDF process ran for each category and sentiment. Gaussian-like distribution are present in almost all the category. For the critique category, the most frequent words, meaning the the highest TF-IDF values, are: vaccines get and people. From these values discussion about vaccinations, the new Omicron variant seems to be the leading discussion topic. For the informative category, analyzed words such as case, new, variant, vaccine and etc. suggest that most informative tweets mentions case numbers, the novel Omicron variant, and information about vaccinations. For the informative questions, the words vaccine, test and curious suggests that citizens are seeking information about vaccines and test on Twitter.

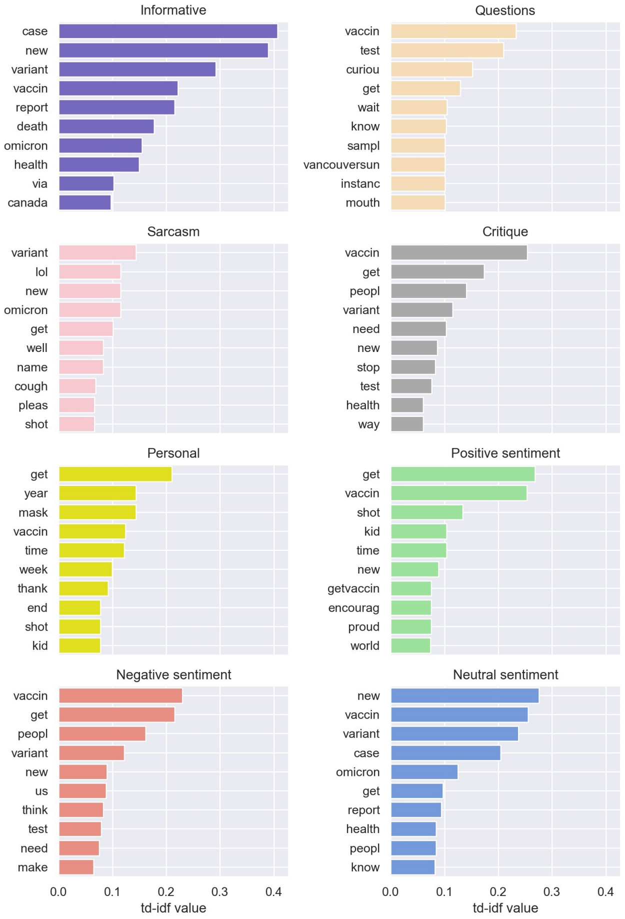

Figure 2: Representation of the proportion of the most popular words in each category. Each category and sentiment is presented with the top 10 words by TF-IDF.

Discussion

The fact the largest portion of tweets were negative critiques reflects that a major focal point of COVID discussion on Twitter in Canada revolves around negative views associated with COVID-related decisions made by the Canadian government and politicians regarding sanitary measures and vaccine mandates. Many Twitter users express hesitancy regarding getting COVID vaccinations especially due to the possible health implications, which is reflected by the prevalence of words such as ”vaccine” in the predominantly negative critiques (Figure 2, Critique). As the critique category makes most of the negative sentiment tweets, TF-IDF analysis suggests that most hesitancy is related to insufficient testing of the vaccines and accelerated delivery of vaccines to the market. This means that the fast vaccine development, which is required during a pandemic, is also the main trigger of vaccine hesitancy and concern for personal health. Also, abundance in the word ”stop”, ”need”, and ”new” also suggests that people criticize and suggest actions to be taken against COVID.

Vaccines seem to be a polarizing topic as they also appear in the most prevalent terms for the positive category. As the data was collected during the time that vaccinations became available for kids between the ages of 5-11, many positive sentiment tweets also mentioned administering COVID vaccinations to ”kids”.

A large proportion of the discourse on Twitter originates from informative posts that do not convey an emotionally charged message. Most of these tweets were news articles reporting the number of new ”cases” and ”death” rates across ”Canada”. As data collection coincided with emergent reports of the new ”Omicron” ”variant”, many tweets were reporting about this new variant. This trend reflects how news outlets are moving to social media as the general public is getting their news from sources like Twitter.

There seems to be confusion about COVID ”vaccines” and ”testing” as reflected by the prevalence of these terms in the questions category. People may be confused about whether they should ”wait” to ”get” their vaccines.

The sarcastic posts mostly revolved around the ”new” ”Omicron” ”variant” and its ”name”, which may sound foreign to most people. The fact that the Omicron appears to be prevalent in many of the categories stems from the fact that the variant had just been discovered around the time when the posts were sampled.

Most tweets were neutral, owing to the fact that the majority of informative tweets refrained from taking any position regarding the pandemic, and were posted by big organization rather than individual users. Conversely, the majority of critique tweets were published by individual users and mostly held a negative sentiment in discussing COVID vaccinations. Therefore, vaccine hesitancy might stem from skepticism.

The largest proportion of positive posts could be found in the ”Personal” category. Many individuals use Twitter to share encouraging stories of ”getting” ”vaccinated” and are encouraging people to wear ”masks”. They may be expressing gratitude and hoping that COVID may come to an ”end”. However, equally many negatives stories related to similar topics.

In summary, the goal of this study was to characterize and classify the discussions about COVID in Canadian tweets and to study the general response of Canada’s population to the pandemic and the vaccination campaign. Results show that major discussions include mostly neutral informative posts and negative critiques about vaccines and sanitary measures. It also shows that Canadian response to the pandemic is mostly negative.

Limitations in this study includes single instead of double annotation process. Further work could link this study with applied government measures, or run comparative quantitative analysis on other social media platforms such as Facebook. Furthermore, interesting conclusions could be drawn from analyzing bi-grams and tri-grams instead of single terms in future studies.

Contribution Statement

All members contributed equally to planning the methods and evaluating the results of this study and were involved in every part of the study. All members labelled all tweets in this study independently, and contributed to writing the report. Alon Shapiro took charge of gathering the tweets used for this study and visualizing all the results. Luca Weishaupt took a lead position in designing the Twitter scraper, the data annotation tool and TF-IDF script. Hugo Baraer was in charge of evaluating and merging the raw annotations, and supervised the final report format.

References

Crabu, S.; Giardullo, P.; Sciandra, A.; and Neresini, F. 2021. Politics overwhelms science in the Covid-19 pandemic: Evidence from the whole coverage of the Italian quality newspapers. *PLOS ONE*, 16(5): e0252034. Publisher: Public Library of Science.

Haleem, A.; Javaid, M.; and Vaishya, R. 2020. Effects of COVID-19 pandemic in daily life. *Current Medicine Research and Practice*, 10(2): 78–79.

Santoveña-Casal, S.; Gil-Quintana, J.; and Ramos, L. 2021. Digital citizens' feelings in national #Covid 19 campaigns in Spain. *Heliyon*, 7(10): e08112.