COVID-19 in Canadian social media

Written by McGill University students
Tuguldur Tuvshinbayar*,
Tuvshin Ganbold*,
Yordan Radev*
McGill University Computer Science department
3480 Rue University
Montréal, QC H3A 2A7

Introduction

A fairly negative sentiment and the tf-idf scores of the top 10 words about 8 curated topics concerning COVID at various levels of engagement are calculated using 1000 tweets, filtered through 18966, for location and keywords. Vaccination sentiment is the most engaging topic with 156 tweets but is shockingly negative with -71.14% total relative sentiment and 77.61% of its tweets having negative sentiment while the only topic with a positive sentiment is joking tweets with 20.21% total relative sentiment and 43.62% of its tweets having positive sentiment. The frequency of tweets collected time encodes important events such as the come to prevalence of the omicron variant which can be seen in Figure 5. The frequency of tweets about vaccine effects and symptoms compared to the one about the COVID effects and symptoms demonstrates the disparity in at least the sharing of such events on Twitter as there are a lot more COVID symptoms related tweets.

Data

Using the official Twitter API, version 1.1, 18966 tweets were collected over the span of 3 days, in Eastern Standard Time. The tweets are collected by a script that runs between noon and 10 PM every hour to cover a longer period that should be more representative of that day. From the streamed tweets we keep the ones that were geolocated within a bounding box shown in Figure 6. We also filter the streamed tweets for a set of keywords that need to be present to make the tweet eligible for the data set (namely covid, AstraZeneca, vaccination, Pfizer and Moderna). These criteria leave us with a rendered down data set that relates to COVID-19 and the vaccination effort for it with a temporal bias for the days of collection, namely November 26th 2021 until November 28th 2021, and a spatial bias, namely the entirety of Canada and the Northernmost states of the United States of America. The API restricted us from being more granular, as to focus only on the greater Montreal area or to the country. The goal of this study being to identify overall trends, we believe the final data set is representative of the Canadian sentiment and engagement around COVID-19.

*These authors contributed equally.

Methods

The final data set of 1000 tweets are cleaned by using regular expressions to remove Twitter mentions (e.g. @elonmusk) as we deemed them irrelevant to the tf-idf and sentiment analysis to be performed later on. Similarly English stop words are discarded to remove the common words that would dominate the tf-idf scores and would not contribute to the sentiment analysis.

In order to proceed, a random sample of the data set was read to identify 8 classes with which the entirety of the data set can be annotated. Figure 1 shows the list for quick referencing through the paper but here is the list and justification in greater detail.

1. **Class 1 :** Reading the sample tweets it became apparent that people took to Twitter to voice their opinion on the vaccination process, vaccines for COVID-19 and a plethora of vaccination sentiment related opinions.
2. **Class 2 :** It also became clear that there is a clear distinction between both proponents and detractors of the vaccination efforts and those of the booster shots. The booster class aims to capture any sentiment or opinion for its class.
3. **Class 3 :** Twitter is seemingly used by both news organisations and individuals to spread information in the form of short descriptive or instructive tweets that are distinct from anecdotes and opinions.
4. **Class 4 :** It became apparent that some people were using the keywords we filtered for in a manner not related to COVID-19 and its vaccination efforts. Examples include businesses using COVID for promotions, individuals mentioning COVID in a tweet about another topic or many other irrelevant to the goal of this study topics.
5. **Class 5 :** Vaccination side-effects, success or failure stories, anecdotes or any tweet related to a specific vaccination administration is obviously a part of what was commonly communicated in our data set.
6. **Class 6 :** Similarly, COVID symptoms, effects and anecdotes are the other side of Class 5.
7. **Class 7 :** Jokes, memes and other media meant mainly as humorous or inflammatory was prevalent enough to be distinguished in this list.

8. **Class 8 :** By a pure coincidence we captured the first days of the omicron variant which resulted in tweets related to variants specifically about variants in general.

Using the above 8 classes the data set was annotated for both sentiment and classification. Each tweet was assigned a value of 1 through 8 designated for membership to one of the 8 classes. Similarly, each tweet was given one of three values to express the sentiment of the tweet, namely -1 for negative sentiment, 0 for neutral sentiment and 1 for positive sentiment. The sentiment annotation is done in a message agnostic manner, meaning both a "pro vaxer" and "anti vaxer" tweet that contains negative language or obvious negative intent ("hope you all die") are considered negative. The data set is annotated by a single person to keep a consistent annotation and because the limited number of tweets.

The annotated and cleaned data set is then tf-idf scored and basic statistics are run on the sentiment and engagement such as :

- **The engagement :** defined as the number of tweets classified with a given class.
- **The total sentiment :** defined as the sum of the negative, neutral and positive sentiments for a given class.
- **The relative sentiment :** defined as the percent of engagement categorized as either negative, neutral or positive over the total engagement for a given class (i.e. 77.61% of tweets being negative for Class 1).
- **The total relative sentiment :** defined as the total sentiment over the engagement of a given class as a percent. This shows the overall sentiment for that class.

tf-idf scores were calculated total of 9 times, one to cover across all the classes, and one for each of the 8 classes. As for the computation of tf-idf scores:

N = total number of tweets

N_w = number of tweets that use the word w across all 1000 tweets

$$\text{tf} = n_{w,i}$$

$$\text{idf} = log_{10}\left(\frac{N}{N_w}\right)$$

$$\text{tf-idf} = \text{tf} * \text{idf}$$

Results

The results in Figure 4 show an uneven distribution amongst the 8 classes as the engagement with certain topics is much stronger than others. Classes 4,1 and 6 are the most engaged with 360,201 and 111 tweets respectively while Classes 2 and 5 are the least engaged in with only 32 and 37 tweets respectively.

As Figure 2 shows, a considerable majority of tweets are negative which is compounded by the fact that the positively categorized tweets are outnumbered exactly 4 to 1.

The negativity of specific classes is less subtle as is the example of Class 1, vaccine sentiment, whose tweets are categorized as 77.61% negative and only 6.47% positive (Figure 4).

The neutral equivalent of this would be Class 3, news and facts, whose tweets are overwhelmingly neutral at 78.65% and have the lowest percentage of negative tweets at 15.73%.

The only Class to have a positive total relative sentiment is Class 7 who is almost equally distributed between jokes of positive, neutral and negative sentiment with positive ones having the highest relative sentiment at 43.62%.

The case of Class 8, variants, is curious as it is the class with the lowest positive relative sentiment at only 5.26% yet it is not as lopsided as Classes 1,5 or 6, but rather split between a neutral and negative sentiment.

The results of the tf-idf scoring, found in Figure 3. Certain words such as "covid" and "vaccine" were ranked among the top 10 tf-idf scored words across most of the classes as expected given the keywords used in the collection process. However, aside from that we see variations that correspond to the topic of each of the classes. Which is shown most clearly in the case of class 8, we see keywords such as "omicron", "variant" and "africa", words not present elsewhere. As such, these tf-idf scores give us an adequate quantification that is representative of the different word frequency characteristics across different classes.

The temporal dimension of the data set, as visualised in Figure 5, allows for a couple of interesting remarks about certain classes. Class 8, variant class in gray in Figure 5, sees a significant uptick in the frequency of tweets once the news of the new Omicron variant starts spreading and is evident in our data set by the cluster of markers and the sparsity of markers before the beginning of November 26th 2021. Another interesting result is the obvious disparity in tweets about Classes 5 (light green) and 6 (yellow). Class 5, vaccine effects, is much more sparse compared to Class 6, COVID effects. Combining this with the results in Figure 4 it suggests that people post negative sentiment-ed tweets much more often about the effects of COVID rather than the effects of the vaccine.

Discussion

The results of the sentiment and engagement analysis shown in Figure 2 and Figure 4 demonstrate a couple of obvious trends. The first trend is the fairly negative sentiment around the topic at hand. As Figure 2 shows a considerable majority of tweets are negative which is compounded by the fact that the positively categorized tweets are outnumbered exactly 4 to 1, creating the impression of overall negativity on the platform even though the negatively categorized tweets are not an overwhelming majority.

Class 3 neutral majority suggests that the news and facts accounts or individuals posting such content make an effort to remain neutral in their delivery of objective facts and simple story lines.

Class 8's split between negative and neutral sentiment suggests that the people have already made their mind in that a new variant is either bad news or at best not certain yet but definitely not good news.

The relevant % in Figure 2 is calculated excluding Class 4 numbers since the Class does not contribute to the topics of concern of this study.

There are several sources of error in this study namely the limited data set, the intersectionality of the classes, annotation methodology, collection methodology and of course the natural variety of human generated content.

The gray plot in Figure 5 explains the presence of the words ”african” and ”africa” in the top 10 highest scoring words in the tf-idf scores found in Figure 3 as the Omicron variant is of African origin.

The count of tweets collected is very low compared to the actual millions of active Canadian Twitter users, meaning the results of this paper are not statistically sound but only a cursory exploration of the current climate meant to inform more specific future studies.

The intersectionality of the classes is a problem that became apparent during the annotation of the data set. The annotator noticed some tweets are related to more than one class in which case he made a judgement call as to which class does the tweet belong more to. A tweet about the negative experience someone had with a vaccine and his opinion about a future booster shot for the Omicron variant is not easily classified.

The annotation methodology is a big source of error in this study for a couple of reasons. Firstly there was a single annotator as to preserve consistency over the limited data set. Naturally, multiple annotators would allow for a much more robust sentiment and classification analysis combating an individual's bias, but due to resource scarcity, it is not practical to have multiple annotators, not to mention for a study at a larger scale. The intrinsic subjectivity of any annotator can only be remedied by a fully diverse and numerous set of annotators who all stick to a stricter guide of at least classification that has much more precise classes.

The collection methodology is the easiest to improve on as it is a simple matter of ensuring a continuous stream of filtered tweets by using the newer version of the Twitter API that allows for more granular Geo specificity. The more data we can gather the clearer and more accurate our analysis would become. Ideally, we would have access to all Canadian specific tweets since the origin of COVID-19 so we can see the progression of trends and compare different variants, vaccines, outbreaks, periods of the year and others.

A byproduct of the natural variety of human generated content is a variety of anomalies which in the case of this study can still be classified but even with classes as broad as ours, there were tweets that were not obviously classifiable. Sometimes there were mixed language tweets. Other times there were tweets that contained a simple meme that can contribute to the sentiment and engagement statistics but would obviously meaningless for the tf-idf analysis. Yet another example would be incoherent tweets that read like bots.

As we can see in Figure 3 the top 10 tf-idf scored words do encapsulate the topic of discussion in each of the classes, which does enforce the validity of our class labelings. One noticeable feature that was not captured in the top 10 list was the presence of obscenities, more prevalent in classes with the highest negative sentiment, which goes to further show the intensity and emotionality involved in those discussions.

When we look at class 6 and class 5 comparatively, we observe more words with negative connotations such as ”died”, ”lost” in class 6. Interpreted this way, it also supports our observation that people's discussion of covid symptoms was more negative than that of vaccine symptoms. In the case of class 3, we have observed that the top 10 words are neutral in tone in correspondence with the topic of class 3 being news and factual communication.

For future iterations of the study, more linguistic characteristics might be captured under an extended analysis inclusive of higher number of words.

Appendix

1. **Class 1 :** Vaccine sentiment in the form of any opinion of vaccines in general and of the vaccination process at large.
2. **Class 2 :** Booster sentiment in the form of any opinion on booster shots in general and of the booster vaccination future.
3. **Class 3 :** Reported news and facts such as new mandates,cases or other factual communication.
4. **Class 4 :** COVID adjacent such as the mentioning of COVID to promote a product or spark a a larger conversation about politics for example.
5. **Class 5 :** Vaccine symptoms and effects related.
6. **Class 6 :** COVID symptoms and effects related.
7. **Class 7 :** Jokes, memes and other media meant mainly as humorous
8. **Class 8 :** Variants of COVID-19 talk

Figure 1: The 8 classes chosen to annotate the data set and their more detailed description.

Category	Overall %	Relevant %
negative	52.4	55.94
neutral	36	30.16
positive	11.6	13.90

Figure 2: Percent of the full data set of tweets per category with and without Class 4, where the exclusion of Class 4 is reflected in the Relevant % column.

Class	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10
Global	vaccination	covid	pfizer	new	variant	vaccine	people	moderna	covid19	get
tf-idf	128.41	118.85	99.43	96.52	80.12	79.38	70.91	66.77	60.60	58.76
Class 1	pfizer	vaccination	vaccine	moderna	vaccines	vaccinated	vax	covid	get	mass
tf-idf	44.20	40.18	38.53	21.83	19.75	17.82	17.63	16.00	13.66	13.33
Class 2	booster	moderna	pfizer	got	boosted	yesterday	boosters	already	shot	get
tf-idf	30.75	21.83	10.04	9.94	7.57	7.56	6.90	6.63	5.76	5.47
Class 3	vaccination	covid19	cases	new	vaccine	pfizer	rate	vaccinations	says	regulator
tf-idf	25.21	24.24	19.44	16.43	16.34	14.06	12.96	12.27	10.94	9.60
Class 4	covid	vaccination	people	cards	checking	without	would	believe	year	amp
tf-idf	49.80	39.39	32.35	20.0	19.59	19.19	19.10	17.63	17.62	17.25
Class 5	vaccination	pfizer	first	dose	shots	jab	vaccine	spike	moderna	effect
tf-idf	8.66	8.03	6.71	6.29	6.0	5.87	5.84	5.40	5.13	5.04
Class 6	covid	died	last	getting	year	today	people	know	lost	back
tf-idf	18.52	16.97	15.49	13.10	12.82	11.60	11.20	10.86	10.0	9.94
Class 7	covid	new	like	pfizer	let	back	get	one	right	need
tf-idf	13.30	11.29	9.56	9.04	8.85	8.29	8.20	8.20	7.83	7.41
Class 8	variant	new	omicron	south	african	covidvariant	africa	strain	covid	already
tf-idf	59.22	46.21	27.20	14.37	14.0	13.44	12.55	11.52	11.33	9.54

Figure 3: The TF-IDF scores for the top 10 words globally and for all 8 classes.

Class	Engagement	Sentiment	Relative Sentiment [%]	Class	Engagement	Sentiment	Relative Sentiment [%]
Class 1	201	-143	**-71.14**	Class 5	37	-22	-59.46
negative	156	-156	**77.61**	negative	26	-26	70.27
neutral	32	0	15.92	neutral	7	0	18.92
positive	13	13	6.47	positive	4	4	10.81
Class 2	32	-12	-37.5	Class 6	111	-65	-58.56
negative	18	-18	56.25	negative	81	-81	72.97
neutral	8	0	25	neutral	14	0	**12.61**
positive	6	6	18.75	positive	16	16	14.41
Class 3	89	-9	-10.11	Class 7	94	19	20.21
negative	14	-14	**15.73**	negative	22	-22	23.4
neutral	70	0	**78.65**	neutral	31	0	32.98
positive	5	5	5.62	positive	41	41	**43.62**
Class 4	360	-139	-38.61	Class 8	76	-37	-48.68
negative	166	-166	46.11	negative	41	-41	53.95
neutral	167	0	46.39	neutral	31	0	40.79
positive	27	27	7.5	positive	4	4	**5.26**

Figure 4: Engagement, in terms of the number of tweets per class, sentiment, in terms of total sentiment and a breakdown of all three categories and relative sentiment, in terms of a percentage of sentiment per engagement are compiled and calculated to reveal trends. The maximum and minimum values for relative sentiment are in bold for all three categories. The total sentiment is a simple addition of all three categories' sentiment and the relative total sentiment is the percent of total sentiment per engagement.

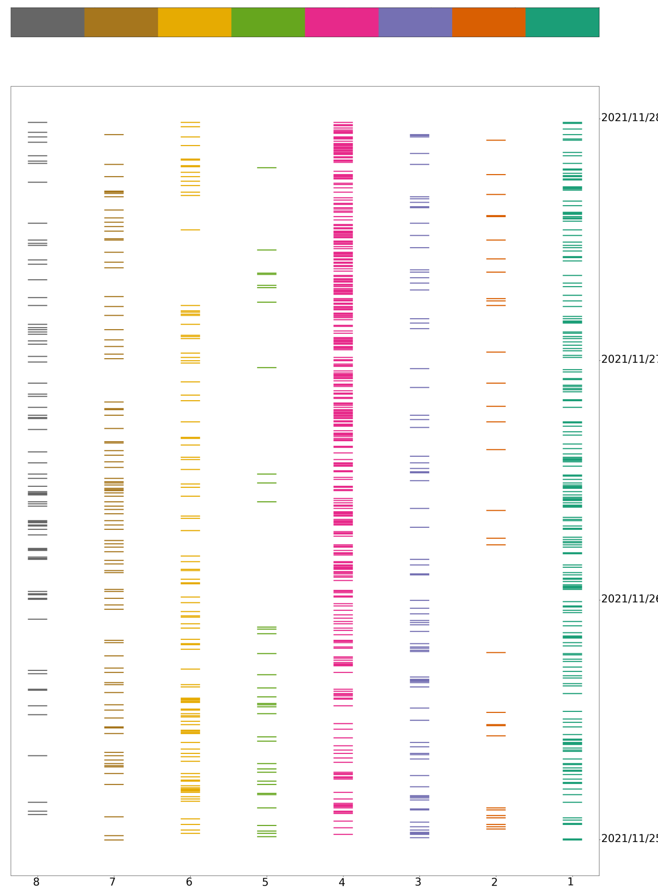

Figure 5: Time series for tweets, per class, for all 8 classes. The time encoded information about certain classes is visualized by noticing the frequency across the 3 days of collection. Class 8 (gray), namely variant related tweets, shows how after the omicron variant appeared the frequency of tweets about variants increased sharply as opposed to the steady inflow of Classes 1,3,4 and 7. It is also interesting to notice the difference in frequency of classes 5 and 6, namely vaccine vs COVID effects, indicating people suffer a lot more from COVID symptoms rather than vaccine effects.

Figure 6: The bounding box (-140.99778, 41.6751050889, -52.6480987209, 83.23324) used to collect tweets.