Comp 598 Final Project

Analysis of Discussions Around COVID in Canadian Social Media

Shawn Hu (ID:260901823), Shanzid Shaiham (ID: 260913075), Doruk Taktakoglu (ID: 260909243)

Introduction.

This paper focuses on the discussions currently happening around COVID in Canadian social media and analyses the main topics discussed around COVID, the relative engagement to these topics, and social media users' sentiments on the main topics discussed around COVID. The analysis is based on the COVID-related tweets posted by Twitter users in Canada in a three-day window.

By looking at the engagement levels of the tweets we have collected, it can be inferred that the most engagement is around the topics of Vaccines and Politics. From analysis, we can conclude that the overall sentiment around COVID-related subjects discussed on Twitter is either Positive or Neutral. In our analysis, we discovered that some of the most frequently discussed topics are news about vaccines, COVID reports, and people's personal experiences and thoughts on the topic COVID.

Data

The data set consists of COVID-related tweets posted in the period 15-17 November 2021. COVID-related tweets are collected for the data set by selecting the tweets that contain the following case-insensitive keywords: covid, vaccination, vaccine, pfizer, pfizer-biontech, moderna, and astrazeneca.

For each day in the data collection period, the latest 400 COVID-related tweets are collected. For each tweet, the time the tweet had been posted at, the location the tweet was posted from, the number of likes and retweets the tweet got, and the text content of the tweet was recorded.

The data collection process is completed after selecting approximately 333 unique tweets randomly from the 400 tweets collected for each day totaling 1,000 tweets. The entire dataset consists of tweets from 534 unique Canadian Twitter users. Most of these tweets were posted by users in Calgary, Edmonton, Saskatchewan, and neighboring regions.

The data collection process is followed by the data annotation process which consists of two parts: categorizing for being a positive, negative, or neutral sentiment and a topic. The annotation process yielded 8 distinct topics. Of the 1000 annotated tweets we have 261 positive sentiment tweets, 523 neutral tweets, and 216 negative tweets.

Data Collection and Parsing Methods

For this analysis, Twitter API version 1.1 was used to collect all recent tweet data from Canada.

Why v1.1 and not the latest v2.0 API?

API v1.1 allows free location-based querying of recent tweets. To get the same functionality in v2.0, a paid subscription is required. Since the project initially requested tweet data from Canada only, our script had already utilized API v1.1 to meet this requirement.

Collecting Recent Tweets from Canada

Tweets data for this project is collected from the region centered at 56°07'49.4"N 106°20'48.5"W (North of Saskatchewan – approx. center of Canada), with a radius of 1000km. This area roughly encapsulates the entirety of Canada, including all the major cities and population-dense areas, as well as a few northern regions of the USA.

Due to the nature of the Twitter API and how it sorts the results, a large part of our collected data is from locations closer to the center of the search region.

Data Collection Process

1. Only English tweets containing the following (case-insensitive) keywords were used in the collection process: covid, #covid, vaccination, #vaccination, vaccine, #vaccine, pfizer, #pfizer, pfizer-biontech, #pfizer-biontech, moderna, #moderna, astrazeneca, #astrazeneca.
2. Only original tweets were collected to make annotation and parsing simpler, i.e., no retweets or tweet replies were collected.
3. Tweets' author names were intentionally left out when collecting data to prevent any bias during any of the annotations/analyses stages.
4. After the first round of data collection, we noticed multiple tweets from different accounts containing the same text (possibly due to bot activity). We then modified our collection script to only collect unique tweet texts and ran the collection process again.

Finally, 400 recent tweets were collected from 15, 16, and 17 November. To ensure all tweets had time to gather engagement, tweets from each respective date were collected 6 days after they were posted.

Note: Due to limitations in the Twitter Search API, we are not allowed to collect tweets more than a week old.

Data Annotation Process

Each member of the group analyzed 50-70 tweets from each date of the collected tweets to find distinct topics. These topics were then shortlisted to the following 8 topics (ordered by highest to lowest *annotation priority*(1)): Vaccine, COVID status, Medical resources, Travel, Politics, Suggestions, Anecdotes/opinions, Unrelated. Tweets that

(1) annotation priority: indicates the priority by which a tweet should be annotated when the tweet's text can match multiple topics. The topic with the highest priority is chosen. E.g. if a tweet can be classified to be in both Vaccine and Politics, Vaccine topic is chosen for the annotation.

were vague in nature were discarded and randomly replaced with better samples from the original dataset of 400 tweets for that day.

After determining the topics for the data annotation process, we created an *annotation priority* for the tweets. Albeit we've selected our topics in a way to ensure that there are no topic overlaps, we made topic Vaccine have the highest annotation priority to make sure that we annotate every Vaccine-related tweet correctly, so we can serve the non-profit organization the most comprehensive and sufficient data analysis possible.

Note: The process of Topic creation consisted of several discussions and revisions which aimed to reduce vagueness, topic overlaps, and subjectivity in the annotation process. Annotations and their Topic definitions and priorities were progressively modified to best suit the specific data we have. For example, initially, we came up with the topic Travel which only consisted of border restrictions, however, during the first annotation process we were unable to place the tweets about social life restrictions (e.g. cancellation of events) under a topic, so we modified the topic Travel to Travel and Restrictions and repeated the data annotation process with the newly defined topics.

During the annotation process, we first check whether a tweet belongs to the topic Vaccine. If it does not belong to the topic Vaccine, we check whether it belongs to the Covid Status topic. If it does not belong to the Covid Status topic, we check whether it belongs to the topic Travel, and if not we continue by checking for other topics in the annotation priority list respectively. This ensured we annotated the most relevant topics for our analysis first and helped reduce subjectivity when annotating.

The sentiment was also analyzed for each tweet to be positive/neutral/negative based on what the text in the tweet was conveying with respect to the COVID or Vaccine topics discussed in this analysis.

Examples -

Negative: "*COVID ruined my summer travel plans*"

Neutral: "*Vaccination rates increased 1.3% this month*"

Positive: "*Anti-vaxxers need to get educated*". Note: Cases like these are where a tweet itself may express a negative emotion, but the information it conveys w.r.t. our analysis is positive (here, the author is in favor of vaccinations), so we choose to annotate such tweets as positive sentiments.

Each member of the group annotated approximately 333 tweets with the topics and sentiments described above, totaling 1000 tweets.

Results

The following are the definitions of the topics chosen for our annotation:

- **Vaccine**: Factual tweets regarding COVID vaccines, vaccination rates, side effects, vaccine passports, and efficacy reports.
- **COVID status**: Factual tweets regarding COVID infection reports, COVID status news, and scientific reports.
- **Medical resources**: Related to medical equipment and resources, supply chain issues, import-export of medical supplies. E.g., masks, respirators, hospital capacity, doctor availability, COVID test kits.
- **Travel**: Related to travel restrictions or event cancellations due to COVID. E.g., border closing, travel bans, canceled events, event capacity limitations.
- **Politics**: Statements from politicians, news reports on politics, and political opinions.
- **Suggestions**: Advice or suggestions offered, e.g., to help prevent the spread of COVID, and raise awareness, use of proper medical equipment (like masks), and advice to reduce chances of transmissions by air.
- **Anecdotes/opinions**: All other non-factual debates, arguments, and personal experiences about COVID of vaccines.
- **Unrelated**: Not constructive to our analysis, e.g., COVID used a metaphor or using COVID as a time period, and awards/recognitions.

Overview

The top 3 discussed topics are Vaccine, Anecdotes, and COVID status with percentages of 24.7%, 18%, and 16.8% respectively. The least 3 discussed topics are suggestions, travel, and Medical Resources with percentages of 2.8%, 3.9%, and 7.9%.

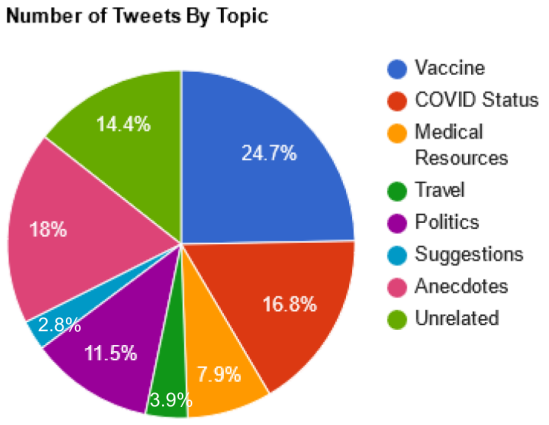

Figure 1: A pie chart showing the percentage of tweets by each topic

Under the topic Vaccine, there is the highest number of positive tweets while the topic with the highest number of neutral tweets is COVID status. The topics with a high

number of negative tweets are Anecdotes, Politics, and Vaccine.

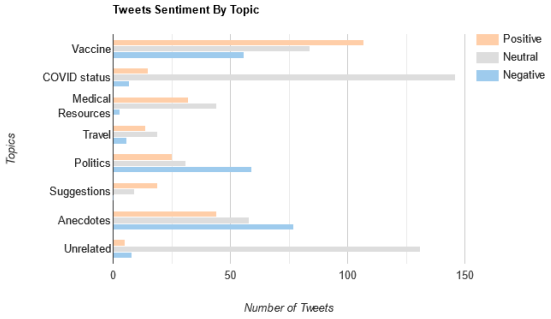

Figure 2: A bar graph showing the number of tweets for each sentiment by topic

Vaccine

The tweets under this topic have a focus on different kinds of vaccines, their potential side effects and the policy related to vaccines, and how people think about it. One of the most mentioned words in this topic is “mRNA” with the highest tf-idf score. While 43.3% of the tweets on this topic have positive sentiments, 34.0% have neutral, and 22.7% have negative sentiments.

The top 10 words with the best tf-idf scores are: "mrna", "code", "anti", "arm", "employees", "immune", "mild", "community", "child", "injured".

COVID Status

The tweets under this topic are mainly news and reports related to COVID cases. The two words with the highest tf-idf scores, "zone" and "central", are used a lot while reporting the COVID cases and situations for location. There are a lot of daily reports on active, newly found and recovery cases thus we have words like "active", "total", "positivity" and "recoveries" on the list. While 8.9% of the tweets on this topic have positive sentiments, 86.9% have neutral, and 4.2% have negative sentiments.

The top 10 words with the best tf-idf scores are: "zone", "central", "active", "mountain", "total", "act", "nw", "positivity", "found", "recoveries".

Medical resources

The tweets under this topic mainly discuss medical resources problems related to COVID. While 40.5% of the tweets on this topic have positive sentiments, 55.7% have neutral, and 4.8% have negative sentiments.

The top 10 words with the best tf-idf scores are: "hinshaw", "approve", "approval", "benefit", "pharmacies", "dr", "weekend", "included", "eligibility", "spectrum".

Travel

The tweets under this topic cover from canceled events to travel restrictions. People talk about how restrictions apply to trips, what games are canceled and shows and movies got postponed due to COVID. While 35.9% of the tweets on this topic have positive sentiments, 48.7% have neutral, and 15.4% have negative sentiments.

The top 10 words with the best tf-idf scores are: "trips", "play", "activity", "games", "marvel", "recast", "super", "checked", "qualifying", "elks".

Politics

The tweets under this topic focus on any news related to politics and people's opinions on politics related to COVID. According to the tf-idf scores, most words are related to the political party such as "council", "party" and "conservative" while others are related to economics such as "cost" and "finance". The word "failure" also has a very high tf-idf score which may be due to the fact that a lot of people are dissatisfied with the performance of the government during the COVID era. While 21.7% of the tweets on this topic have positive sentiments, 27.0% have neutral, and 51.3% have negative sentiments.

The top 10 words with the best tf-idf scores are: "cost", "council", "party", "repeal", "finance", "wotherspoon", "failures", "meeting", "wong", "conservative".

Suggestions

The tweets that fall under this topic make COVID-related suggestions to people and the government. While 67.9% of the tweets on this topic have positive sentiments, 32.1% have neutral, and none of them have negative sentiments.

The top 10 words with the best tf-idf scores are: "duration", "improve", "classes", "bathroom", "mask", "winter", "leaders", "ventilation", "vigilant", "using".

Anecdotes

The tweets under this category primarily consist of non-factual or opinionated discussions, personal experiences with COVID or vaccines, or debates/arguments with other Twitter users. While 24.4% of the tweets on this topic have positive sentiments, 32.2% have neutral, and 42.8% have negative sentiments.

The top 10 words with the best tf-idf scores are: “ha”, “talking”, “mom”, “trying”, “boss”, “listening”, “examples”, “sick”, “feel”, and “im”.

Topic	Total likes	Average likes		
		Positive tweets	Neutral tweets	Negative tweets
Vaccine	4648	6.2	3.7	18.4
COVID Status	783	5.6	4.0	16.4
Medical resources	305	5.4	3.0	0
Travel	43	1.5	0.1	0.6
Politics	2476	70.6	4.7	9.6
Suggestions	700	24.3	26.4	0
Anecdotes	1885	7.7	12.6	10.6
Entire dataset	11338	13.4	5.4	23.3

Figure 3: Tweets' Engagement By Topic

Discussion

The Overall Response to Vaccines, COVID, and the Pandemic

In regards to vaccines, our data indicate that Twitter users posted the largest number of positive tweets regarding Vaccines - over 10% of our entire dataset consisted of positive Tweets around the Vaccine topic. At the same time, it is important to note that the negative tweets on this topic received significantly greater engagement than the positive tweets, as evidenced in Figure 3. We can generalize this as a form of cautious optimism around vaccines, where the general sentiment is positive, however, any doubts or negativity is also being engaged with significantly.

People tweeted about different kinds of vaccines and their side effects or related injuries, especially in children and immunocompromised people. They are also concerned with its potential side effects in vulnerable groups such as children. People are concerned about how vaccine policy affects the work and community, and dislike compulsory COVID vaccines. There are also many tweets talking about "anti-vaccine" as well, which is the word with the third-highest tf-idf score. While a lot of tweets talk about how vaccines help to develop immunity, a great number of them are about there are serious side effects. The word "Hinshaw" is one of the words with a high tf-idf score because Deena Hinshaw is the Chief Medical Officer of Health for the province of Alberta and there is a lot of news about the medical resources during the days when these tweets are posted. The tweets also have a focus on the shortage of medical resources due to COVID. Words like "mask" and "ventilation" are mentioned a lot because these tweets try to remind people of their importance and make suggestions accordingly under this topic.

Addressing the High Engagement On Negative Tweets

In the entire dataset consisting of 1,000 tweets, 26.2% of tweets have positive sentiments, and 21.6% of the tweets have negative sentiments. Although only 21.6% of the tweets have negative sentiments, negative tweets receive an average of 23 likes whereas positive tweets receive only 13 likes, as seen in Figure 3. Although negative sentiment tweets do not appear as much as positive or neutral sentiment tweets, when such negative tweets get posted, they invoke a stronger response and a bigger engagement. This may be caused by the fact that those Twitter users, who engage with such negative tweets by liking or retweeting them, actually relate to those tweets. They may have experienced the same things as mentioned in the tweet that they are yet to tweet about. Another reason for this high engagement for the negative tweets may be caused by these negative tweets creating a shock or an emotional response for the users who end up sharing the tweets, which in turn helps boost engagement on these tweets.

With respect to the Politics topic around COVID, 51.3% of the tweets annotated have negative sentiments, which is supported by the fact that for political tweets, "failure" has a relatively high tf-idf score that shows dissatisfaction of the community with the government.

However, while positive sentiment political tweets have 70.6 likes per tweet, negative sentiment tweets have 9.6 likes per tweet, and neutral sentiment tweets have 4.7 likes per tweet.

This mismatch in the number of tweets vs their relative engagements may also be explained by the fact that not every Twitter user has a similar number of followers, similar credibilities, or similar influences on the Twitter platform (e.g. controversial tweets from influential people may be generating high engagements, which in turn get promoted to even more users, creating a feedback loop). In Politics, for instance, this high like per tweet ratio for positive sentiment tweets may be caused by the fact that some tweets in this category are posted by politicians or public figures, who have a lot of followers, thus a lot of credibility and influence on the platform. So it is reasonable to assume that such public figures' or politicians' tweets are expected to have more engagement compared to tweets posted by non-famous Twitter users. This can apply to any of the other topics we have as well and can help explain the high engagements on negative tweets.

Addressing the High Number of Tweets Under the Topics Covid Status and Vaccine

There are a large number of tweets about daily reports and news on covid status and cases which contributes a lot to the high number of tweets under topic Covid status. As of topic Vaccine, there is a lot of new information every day on either the vaccine itself or any policy related to vaccines such as QR code therefore there is a huge number of discussions and debates about vaccines on Twitter which makes it a popular topic to tweet about. Another reason is that Vaccine and Covid status are of the highest priority among the topics thus we tend to annotate tweets with them if related instead of other topics.

The characteristics of the tweets under topic Covid status are that they are neutral in terms of sentiment because most of these tweets are news or reports on updates of covid status, which is the major reason that Covid status has such a high number of neutral tweets.

Outliers and Vocal Minorities

For tweets categorized as Anecdotes, our data indicate a strong skew towards negative tweets, as well high engagement in these negative tweets. These tweets generally discuss non-factual, or opinionated statements around COVID, and were posted by 66 unique users (which accounts for 12% of the 534 unique users we have in our whole dataset). This shows that we have a small minority of users who strongly feel negatively about COVID and they also receive significant engagement for their views.

Suggestions & Improvements

Certain events or news during the data collection period will lead to a relatively high number of tweets under related topics which cause biased data, especially with a 3-day data collection window. Instead of collecting data from a 3-day window, we could consider collecting data over a longer period therefore any events or news happening during the data collecting period will contribute less to biased data, which leads to a more meaningful result.

In the method section, we mentioned using one location, which is the center of Canada, as a parameter to filter data. Instead, we could consider dividing 1000 tweets into different batches and using locations of different cities for each batch during the data collection so that our data set would cover data from specific cities.

Group Member contributions

Shanzid Shaiham

- Set up Twitter Developer account and GitHub organization for code sharing
- Created python script for collecting and parsing tweets using the Twitter API
- Organized weekly sync meetings to discuss updates and track progress
- Annotated tweet data with topics and sentiments (333 tweets out of 1000)

Doruk Taktakoglu

- Created a Python script for analyzing the engagement levels (likes and retweets) of tweets according to their topics and sentiments
- Created a Python script for analyzing the number of tweets for each region.
- Annotated tweet data with topics and sentiments (334 tweets out of 1000)

Shawn Hu

- Add a git repository data-analysis
- Created a python script for calculating the tf-idf scores and compiling ten words for each topic based on the ranking
- Annotated tweet data with topics and sentiments (333 tweets out of 1000)