Discussion Analysis regarding COVID-19 Vaccines

Jessica Pizzuco, Nuri Ege Zararsiz, Aly Saleh

COMP598 - McGill University
jessica.pizzuco@mail.mcgill.ca
nuri.zararsiz@mail.mcgill.ca
aly.saleh@mail.mcgill.ca

Introduction

The COVID-19 pandemic is currently one of the major topics discussed over social media. In an effort to understand the discussion currently surrounding COVID-19, we have collected and analyzed data from 1000 tweets in order to gain insight on vaccine hesitancy, the major topics discussed around the subject of COVID, and the sentiments attributed to those tweets. This involves conducting an open coding on 200 tweets and finding 7 unique topics, calculating the tf-idf scores of the words in these topics, and annotating all the data in order to categorize the data by sentiments of positive, neutral, or negative. Through various charts and graphs, we see that the majority of tweets discussing COVID-19 restrictions disapprove of them. However, for tweets discussing personal opinions and experiences, around half of them supported vaccination and the other half did not support vaccination. We also discovered that tweets labelled as disapproving of vaccination resulted in a lot more discussion than tweets labeled supporting vaccination as indicated by the higher retweet and favourite count. Furthermore, positive news have high engagement, demonstrating a desire for hope and normalcy. Although filtering by recent tweets rather than ones with the highest engagements, news and media seem to be the most distributed in the dataset, allowing the headlines from various institutions to continue to spark conversation and provoke feedback from the population.

Data

To form the dataset, we used Twitter's API and the official Twitter API library, Tweepy. The main reason for using Tweepy is that it is the most well-established twitter library we could find. To collect tweets, we searched for tweets that contain the following keywords, collected case-insensitive:

- covid (200 tweets)
- vax (200 tweets)
- vaccination (200 tweets)
- moderna (80 tweets)
- astrazeneca (80 tweets)
- pfizer (80 tweets)
- vaccine (160 tweets)

We collected 1000 tweets over the span of three days. We selected three of the most widely distributed COVID-19 vaccine manufacturers in English speaking countries (Moderna, Astrazenca, Pfizer).

The Twitter API returns a lot of information with each tweet, such as the user's username, when their account was created, the description the user has on their Twitter account, etc.. For the purposes of this project, we only extracted the following relevant fields from each tweet in order to easily navigate through the data:

- ID: each tweet is assigned a unique ID. We used this to ensure that tweets don't show up multiple times in the data set.
- Text: the text content of the tweet, which is an essential element for the analysis. We used this to determine the topic and sentiment of the tweet.
- Favourite count: the number of likes on a tweet. We used this to identify tweets that were supported by Twitter users.
- Retweet count: how many times the tweet was retweeted. This variable is important since it indicates the amount of engagement and discussion that is generated from the tweet.

Our data set does not have any retweets or tweet replies. Downloading all the tweets (original and retweets) returned by the search_tweets function, and then filtering out the retweets exhausted our API capacity. To solve this, we investigated the search_tweets function options and discovered there was a way to have the API only return original tweets without any retweets. The data collection process was completed once we ensured the following:

- Retweets were not returned by the API.
- Tweet replies were not returned by the API.

Methods

We decided to collect the data by searching for tweets with the keywords mentioned in the previous section. These keywords were used because in almost all the tweets, they were related to the COVID-19 pandemic. For example, using words like virus was intentionally avoided since we would get tweets that might not be relevant to the COVID-19 pandemic. We collected only 80 tweets for each of the major

vaccine manufacturers (Moderna, Astrazenca, Pfizer), since we did not want bias toward a particular company to skew our data. Additionally, different tenses of verbs were not used since the Twitter API handles this if we used the present tense of a word. For example, using the keyword vaccinate as a keyword for search, tweets that include the words vaccinating, vaccinated, vaccinates are all collected. After collecting the data set for the first time, some of the tweets collected were actually *retweets*, not handled by analyzing the IDs of the tweets since each retweet had its own unique ID. These retweets had to be removed for two reasons:

- If we included retweets in the dataset, there would be duplicate data that our later analysis may favour.
- The aim was to collect 1000 tweets. Including retweets would mean that there are less than 1000 unique tweets in the dataset.

After collecting the dataset without any retweets, we encountered another issue. It seemed that a lot of the tweets lacked any context. Even though the tweets contained one of our chosen keywords, the API was returning tweets that were part of a larger conversation, such as responses. To better understand why this was happening, we learned more about what the search_tweets function returns. We discovered that it also returned tweet replies, which are like comments. These tweet replies are separate from retweets, that is why they were included in the APIs response even if we specified no retweets. Tweet replies lacked context since they were part of a larger conversation that was not captured by the Twitter API, so we had to remove them. Tweets were collected based on recency rather than popularity for two reasons:

- Popular tweets were mainly posted by people with a large number of followers, such as politicians and actors. However, recent tweets includes the tweets that are both popular and recent, and are therefore a better representation of the general population.
- The twitter API did not return enough tweets that are popular for us to have 1000 unique tweets in the 3 day span of data collection.

Once the tweets were collected, we needed to extract and analyze certain aspects of the data. In order to characterize our data, the top 10 words in each category with the highest tf-idf scores were computed. This was done by writing a Python script to compute the following:
`tf-idf`(word, category, wordcounts) = tf(word, category) · idf(word, wordcounts), where tf(word, category) computed the number of times a category contained *word*, and idf(word, wordcounts) computed

$$log\left[\frac{(\text{total number of categories})}{(\text{number of categories containing that word})}\right]$$

Finally, more filtering had to be done when calculating the td-idf scores since there were some issues that arose when trying to run the code on the dataset. The first thing that had to be done was to filter out *stopwords* and to remove symbols in the tweets the were not impotant and were going to skew the results. For instance, while collected the tweets, emojis and certain punctuation and symbols was converted to unicode, so those symbols were removed through the use of various regular expressions, as well as links. Lone numbers were also removed as they were not useful for analysis. Another issue that arose was with the quotation marks in the csv files. Often times there were quotes within the Text field of the tweets, causing the script to break. In order to fix this issue, inner quotes in the tweets were replaced by single quotes using an Emacs macro. While reviewing the results of running the tf-idf implementation, there were a few adjustments that were made. Firstly, there were words that were very unique to some categories but not very meaningful for the purpose of this analysis. For example, certain names of various doctors or various Twitter handles. Therefore we cleaned up the data by restricting to only words that showed up greater than three times, as well as removing Twitter handles before calculation.

Moving on toward the analysis, an open coding was conducted in order to build up a typology. The steps taken in the open coding process were the following:

- Take a sample of data. The sample size was 200 according to the project instructions.
- Examining the sample data and assigning topics to each tweet, without restrictions.
- Re-examining the sample several times and by different members of the group in order to come up with unique and non-overlapping 3-8 topics.

The sample was selected randomly from our data set, eliminating subjectivity in our sample. The first topic assignment process yielded more than 10 topics which included multiple overlapping topics. After two rounds of re-examination conducted by different group members, the number of topics was reduced to 7.
This selection of topics reduced overlaps between tweets. Furthermore, in order to minimize subjectivity and the number of possible errors made by different annotators, an annotation guide was written and closely followed. For each topic, this guide provided an example tweet that belongs to the topic, a second example that does not belong to the topic, and a gray area example where it wasn't directly clear if the tweet belongs to the topic or not, and how to resolve these "gray area" situations. According to the sample tweet guidelines, each tweet had to be annotated according to the sentiment. A plan was devised on how our group approached the sentiment depending on the topic and this plan was included in the annotation guide in detail. Then, all 1000 tweets in our data set were annotated following this guide. For analysis, graphs and charts were created using Matplotlib, as well as generating several json files in an effort to closely interpret various results and calculations. These files contained raw data that was used in order to make tables of relevant findings in the data and provide us with numerical values to calculate statistics based on the sentiment analysis.

Results

After conducting an open coding, 7 unique topics were obtained:

- Announcement
- Restriction/Mandates/Pandemic Related Policies
- Conspiracy
- Politics
- Personal Effects of COVID/Pandemic/Vaccination
- Personal Opinion about COVID/Pandemic/Vaccination
- No Context

In an effort to assess the sentimental response to the pandemic/vaccination, the topics we found must be unique and well-defined. Each topic can be defined as follows:

- Announcement topic gives useful insight on the media coverage of pandemic. This topic is defined to include every announcement/sourced information from public or private institutions, therefore, it was the most common topic we encountered during our annotation process. To stay objective with this topic, sentiment towards news/announcements was measured by the content they covered rather than pro/anti vaccination. For the sentiment to be negative, the content needed to include *objectively* negative statements such as death. In parallel, for the sentiment to be positive, the content needed to include *objectively* positive statements such as recovery from COVID-19. Example: "Pharmacists promoting Covid-19 booster during National Influenza Vaccination Week" is labeled as a neutral announcement.
- Restriction related policies was another big topic in our annotation and analysis because it gave insight from both anti and pro vaccination groups. In our annotation guide, a condition was specified which gave priority to this topic. For example, if a tweet consisted of news information that was about COVID-19 restrictions, this topic was chosen rather than announcement topic. The strategy of sentiment annotation towards this topic was unlike the strategy used for announcement topic because we needed what people thought about the mandates. This *does not* mean that the sentiment analysis was subjective because the sentiment annotation *does not* benefit any group. It merely gives an insight of what people think towards the idea of restrictions. Example: "I really wanted to eat out but all restaurants require vaccination passports and I don't have one, I hate this" is labelled as a negative sentiment.
- Conspiracy topic has the following definition in our guidebook: "Publishing unassisted/non-scientific information, lack of trust in public health institutions and science, personal beliefs that are nonparallel with objective findings". To stay objective with our annotation, we avoided giving personal judgements on this topic. These topics nearly always gave *negative* sentiment toward vaccination. Examples: "So now we will have to wait til 2096 to see what's in Pfizer's poison" is considered a negative sentiment in the conspiracy category.
- Politics was a subject that was mainly covered in news articles related tweets. Priority was given to the announcement topic over politics because we were mainly interested in the sentiment of the public towards the pandemic/vaccination. While politics played a huge role in people's sentiments, when the most coverage was done by media corporations, the idea of prioritizing politics and sentiment toward politically-related announcements was not aligning with our end goal. The tweets that were sent by individuals without any announcements/news were annotated with this topic and the sentiment towards this topic was again, towards restrictions/vaccinations. Example: "The government should do more to protect people from COVID-19" is labelled as negative sentiment under politics.
- Personal Effects/Experiences included the tweets where people mostly talked about their vaccination/COVID/pandemic experiences and the effects of these topics on them. However, restriction related experiences were prioritized to the restrictions topic. This topic gave us mostly balanced answers as people having positive or negative experiences with the vaccine or COVID-19 balanced themselves out. Example: "I am glad I finally got my COVID-19 vaccine!" is a labelled as a positive personal experience.
- Personal Opinion topic included the tweets where people talked about their own opinions and were being sarcastic/making jokes about the pandemic/vaccine/COVID. It was important to annotate this topic with at most attention because subjectivity had to be avoided and it was important to respect all ideas while annotating. We did not annotate people criticizing or commenting about the vaccination as conspiracies unless they were publishing/claiming misinformation. Sentiment was annotated according to people's opinion on the vaccination. Since there were abundance of jokes about the vaccines and COVID-19, there is a considerable amount of neutral tweets. Example: "I think it is good we have access to vaccines to defeat COVID-19" is labelled as a positive opinion.
- Our goal was to avoid the "Other" topic as it did not provide us with relevant information. Although we were successful in that aspect, we were lacking the necessary information to fairly annotate some tweets, therefore we used No Context topic in those cases. There were not many cases under this topic so our analysis still gave us meaningful insight. This topic usually involved buzz words without any meaningful content, and were therefore mainly labelled as neutral in the annotation process.

Using the tweets we had collected in various csv files, we were able to compute the tf-idf scores of the words in each category. The 10 words in each category with the highest tf-idf scores are presented case-insensitive, as the were collected, in the table below:

Announcement	Restriction	Conspiracy	Politics	Effects	Opinion	No Context
astrazeneca	employees	transparency	political	booster	anti	flu
protection	mandate	data	biden	arm	astrazeneca	truth
study	military	pfizer	trump	dose	friend	virus
omicron	nationwide	passports	republican	double	sick	pandemic
pfizer	contractors	clean	support	pfizer	don	news
accused	federal	act	wearing	moderna	mask	vaccination
health	judge	fda	workers	home	effective	vaccine
lethal	required	coming	administration	lost	lecture	covid
fake	requirement	unless	ppl	heart	love	vax
protect	mandatory	countries	mandate	life	mother	update

Table 1: 10 words with highest tf-idf values per category

This table allows us to get a good representation of what the most common unique words were for each topic. With this, we discussed other data to collect and analyze in order to deepen our understanding of how various groups felt toward COVID-19, vaccination, as well as how they expressed these feelings. As well as this, we produced a pie chart that displays the overall percentage of tweets per category:

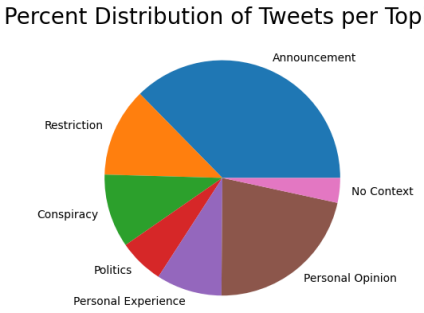

Figure 1: Percentage of tweets making up the total 1000 for each topic

This chart helps to understand how many users were actively talking about specific topics. The percentages for each topic are broken down in the table below:

Topic	Percentage
Announcement	37.34%
Restriction	12.21%
Conspiracy	10.11%
Politics	6.21%
Effects	9.01%
Opinion	21.72%
No Context	3.40%

Table 2: Total percentage contribution per category

Although this was useful is discovering information about which categories were the most active on Twitter, we wanted to analyze the *sentiments* of the users, with respect to the topics chosen. In order to gather this data, we produced a 3-stacked bar graph, that presents the tweet occurrences of each topic, as well as plotting the sentiment that was discovered during the annotation process for these tweets. These findings are presented below:

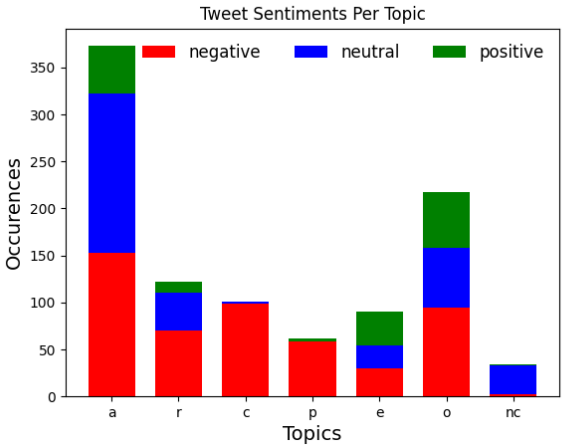

Figure 2: Tweet sentiments per topic with relation to their occurrences

With this graph, we were better able to directly see correlations between tweet occurrences per category, the sentiment behind them, and take a closer look at how certain topics are presented online based on these results. Next we wanted to split our findings into explicitly displaying how these interactions were broken down.

In order to take a closer look at topic engagement, we constructed two graphs. Once representing the average number of *favourites* on a tweet, separated by topic, and the other representing the average number of *retweets* on a tweet, separated again by topic. The graphs that follow, respectively, show the sentiment for the categories, allowing us to determine if there is a relationship between tweet engagements and their sentiment, with respect to their categories:

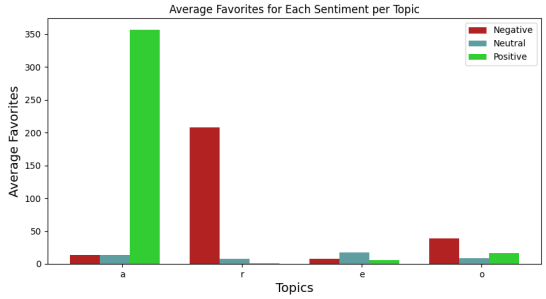

Figure 3: Average favourites for each sentiment per topic

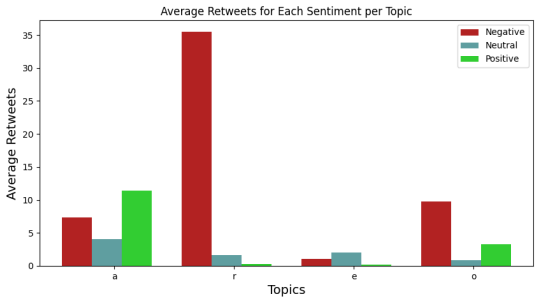

Figure 4: Average retweets for each sentiment per topic

For the tweets under the Restrictions topic, the high average retweet count is due to the restrictions that are labelled as negative. Note that this graph also reveals that the average retweet count for restrictions labelled as negative is 146.50 times higher than the restrictions that are labelled positive.

Another note is that the Announcement topic has a high favourite count. The high average favourite count for announcements can be attributed to announcements labelled as positive.

For the Personal Opinion topic, tweets that are labelled as negative have an average favourite count that is 2.38 times higher than those labelled as positive. Additionally, The average retweet count for negatively labelled tweets is 2.95 higher as seen in Figure 4.

Figure 4 shows that the average retweet count is the highest for Restrictions, which can be attributed to the tweets labelled as Restrictions with negative sentiment (an average of 35.46). On the contrary, retweet count is very low for Restrictions labelled as positive.

Another notable observation is the Personal Opinion section, where the average retweet count for a negative tweet is 2.95 times higher than a Personal Opinion tweet labelled as positive.

It's clearly seen in these graphs that the Announcements topic has the highest number of tweets, as well as the highest number of average favourites for positive tweets. However, it's interesting to note that the Restrictions topic has the highest number of average retweets, and in particular, they are those tagged with having a negative sentiment.

Discussion

From the results found in these graphs, we can combine several sources of information and data to begin to interpret and understand several different things about how people have reacted the COVID-19 pandemic as well as vaccines. Using the graphs, we see that tweets that are labelled as Restrictions and have a negative sentiment have a much higher favourite count and retweet count than Restrictions that are labelled as positive. From this, we can infer two things:

- Since the negative restriction tweets are classified as tweets that don't agree with or disapprove of restriction measure, while positive restriction tweets indicated agreement and support with restriction measure, we can infer that generally, Twitter users support tweets that disapprove of restrictions.
- Along the same lines, tweets that disapprove of restrictions result in a lot of discussion and engagement, as indicated by the high rate of retweets.

Hence we can infer that tweets that disagree with restriction measures have a lot of support and engagement compared to tweets that support restrictions.

We can also see that the Announcements labelled as positive sentiments generate more discussion and support compared to ones labelled as negative, indicated by the higher favourite count and retweet count.

Negatively labelled opinion tweets indicate negative opinions on vaccines. From what we see in the graphs regarding negatively labelled opinion tweets, which have both a higher average retweet count and average favourite count compared to positively labelled retweets, we can infer that negative opinions of the vaccine result in higher engagement and support compared to tweets that support vaccines.

Unsurprisingly, the Announcements category has the highest number of annotations in the dataset, corresponding to 37.34% (Table 2) of the total tweets collected. This is because Twitter has become a platform that millions use to get their news. There were many tweets from public or private news institutions that were shared, quoted, or reworded, therefore resulting in a high percentage of tweets in this category. It makes sense then that the results from Figure 3 show that Announcements have the highest average number of favourites for positive sentiments. People want to hear that the world situation is improving, so when they see these positive tweets from media that is "trusted", the engagements with these topics increase.

Unfortunately, the opposite is also true. We see that Figures 3 and 4 both show that interactions are high for tweets that have a negative sentiment and are discussion restrictions or mandates. This is where discussion and debate occurs. High retweet counts may be due to people spreading awareness of new restrictions and high favourite count may be people who are upset that restrictions weren't put in place earlier, or that they are still getting stricter.

There is insight to be gained from Figure 2, where there is a clear representation of how topics are divided. As discussed, Announcements have the highest number of occurrences, and there is a pretty even distribution of negative and neutral sentiments from those tweets. This does not however mean that there is bias news, since our annotations avoided bias and tagged sentiment for this topic by considering only *objective* good and bad, it means that institutions often times tweeted things involving death, declining resources, inability to maintain public health, etc., while almost rarely speaking of any positive news.

For the most part, the highest sentiment per category was negative. Figure 2 shows that Conspiracy and Politics topic had overwhelmingly negative sentiments. By analyzing the 10 words from each category with the highest tf-idf scores, we can better understand what the main topic of conversation was in the respective categories, and gain insight on how that corresponds to what we have seen regarding tweet en-

gagement/interactions per topic.

The words for each topic with the highest tf-idf scores align well with the other data we found. Table 1 shows that Announcements often mentioned “vaccines”, “studies”, the new Omicron variant, all topics that continue to be prevalent and constantly updated with new information. Restrictions mention “employees”, “mandates”, the military, and those who are either affected by these restrictions or those who are putting them in place. Conspiracy has words that help us understand why the sentiments were so negative for that topic. With words like “transparency”, “data”, and “FDA” taking the top spots. There is clear distrust in a lot of the information that has been transmitted to the population throughout the past year and a half. Politics, as expected, mention the names of several world leaders and political parties, such as Biden, Trump, Republican, administration, etc. Effects mention more personal topics that we mentioned in tweets from people sharing their own personal experiences, with, looking at Figure 2, pretty evenly distributed data. This involves words like “booster”, “arm”, “dose”, referring to people getting vaccinated and sharing the effects it has had on them, whether they were positive or negative. Finally, Opinions mention things that can be seen as controversial by some, with some being in favour and some being against topics of importance such as vaccination, the validity of Covid, wearing a mask, etc. These ideas manifest into words such as “anti”, “effective”, and “mask”, while more personal tweets involve words like “friend”, “mother”, and “love”. The No Context category is filled with buzzwords. This is because they were tweets that did not contain much content or context, but rather several trending words like “flu”, “truth”, “virus”, etc. Overall the combination of all the data collected allows us to learn more about how people dealt with the pandemic, vaccines, and how they engaged with certain information over others. While positive sentiment tweets had the highest number of favourites, negative sentiment tweets got discussions going and were spread around more overall.

Group Member Contributions

Nuri Ege Zararsiz:

- Open Coding
- Building Typology and Annotation Guide
- Data Annotation
- Data filtering
- Chart/Figure Creation
- Methods and Results writing on the report

Aly Saleh:

- Open coding/refining the typology
- Data collection
- Data cleaning and formatting
- Data annotation
- Data analysis and discussion (Retweets and Favourite count)
- Writing data collection section
- Writing part of results and discussion

Jessica Pizzuco:

- Data filtering
- tf-idf implementation
- Data annotation
- Report formatting
- Creating tables
- Writing Introduction, Results, Discussion
- Editing report