COMP 598 Final Report
COVID Discussions in North American Social Media

Mark Bai, Sophia Ju, Nicole Xu
McGill University
hao.bai@mail.mcgill.ca
sophia.ju@mail.mcgill.ca
yue.xu@mail.mcgill.ca

Introduction

Coronavirus disease (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus. This virus can lead to mild to moderate respiratory illness and even though most of the people can recover without requiring special treatment, older people, and those with underlying medical conditions like cardiovascular disease, diabetes, chronic respiratory disease, or cancer are more likely to develop serious illness. (WHO, 2021) (Kraemer et al. 2020). The virus has caused unprecedented change on the entire world since its outbreak in 2019, and as a result many facets of society related to it including vaccination, new restrictions, and politics have become massive areas of discussion. In addition to its impact on public health, COVID-19 has severely affected people's normal lives.

Popular social media applications are an accessible way for people and organizations of all kinds to share their thoughts with the rest of the world. Twitter is a platform for people to communicate and stay connected through the exchange of quick, frequent messages. People post Tweets, which may contain photos, videos, links, and text; it is also one of the most widely used web applications where all messages are discoverable on Twitter search. (New user FAQ, 2021) Therefore, to analyze the influence and the sentiment towards COVID, we developed some statistical methods using Python to extract Tweets from Twitter where we can analyze the data quantitatively.

The overall purpose of our project is to collect and understand the discussions happening around COVID and vaccine hesitancy within the North America region. We aimed to understand the primary topics that Twitter-users are concerned about, as well as their attitudes towards those topics. To do so, we obtained the top 10 most frequently used words in these topics by using TF-IDF scores.

We selected more than 1000 tweets related to COVID within a 4-day range from November 27th to Nov 30th and after filtering and classification, we found out that most of the topics were related to new COVID variants, vaccine and general COVID news and information. A large percentage of the Tweets were either negative comments towards the virus or vaccine or neutral message such as news or explanations. The prevalence of tweets in these categories show the huge impact COVID has made to everyone in their daily lives.

Data

To collect our data, we first formed a simple query consisting of a few keywords a tweet must include when requesting new tweets from the Twitter API. This query filtered for tweets that were NOT retweets, that had the language field set to English, and contained any of the following words: "covid", "covid19", "quarantine", "pandemic", "vaccine", "vaccination". At first glance, this query seemed to result in tweets that were applicable to our analysis. However, upon further inspection, we noticed that some tweets that contained the word "vaccine" or "vaccination" were discussing diseases other than COVID such as the flu or measles.

In order to capture tweets that were specifically discussing COVID, we changed our query slightly so that if a tweet has the word "vaccine" in it, it must also include "covid", "covid19", "pfizer" or "moderna" as a word in the tweet. This way, we could be sure that if a tweet mentioned a vaccine, that the topic was about COVID vaccination and not any other vaccination discussion. Additionally, we also included tweets that mentioned just "pfizer" or "moderna" (case-insensitive) in their text, because these brand names are specific to COVID. Other brand names were not included, as the majority of Canadians have received doses by these two name-brand vaccines (Public Health Agency of Canada 2021). In the end, the exact query that we used to obtain our data is as follows:

> (vaccine (covid OR covid19 OR pfizer OR moderna) OR coronavirus OR covid OR covid19 OR quarantine OR pandemic OR pfizer OR moderna) lang:en -is:retweet

This query was used to collect tweets over three consecutive days, with 350 tweets collected on the first and second day, and 300 more collected on the last day for a total of 1000 tweets.

Once we started the open coding process, we noticed that some tweets had been included in the results of the query that did have any of the keywords in them. However, we did observe that all of them contained at least one url. To

figure out why they were being picked up by the query, we looked up the tweets by their unique ID to find them on the actual Twitter application. From doing this, we found that these tweets all contained a keyword in the expanded url and that this was hidden from us because the query request only returns short urls. For example, the body of the tweet in the request result looked like this:

> @SunnyL723 In June 48hr turnaround was the aim https://t.co/dxDg1tKMnh

but the tweet with the expanded url looked like this:

> @SunnyL723 In June 48hr turnaround was the aim https://www.gov.uk/government/news/uk-surpasses-500000-coronavirus-covid-19-tests-genomically-sequenced

This example tweet ended up being one that is relevant to the COVID discussion, but we only had access to the short url in the request result that we used to code the tweets. Unless we repeated this process for every tweet that this occurred, we would have no context for what the tweet was about. To avoid using these tweets in our analysis, we first debated if we should filter out all tweets that contained urls by changing our query, but decided against this because we felt that it would remove too many relevant tweets. After looking through more of our data, we noted that this situation happened rarely enough that we could manually delete these tweets whenever we came across one. Additionally, there were some tweets that were not in English but had their language field set to English, and we needed to delete these ones as well.

Because we had to manually delete some tweets, we obtained 50 more tweets on the fourth day in order to make up for the ones we would have to remove. We went through the coding process for all 1050 tweets and ended up needing to delete 30 of them, leaving us with 1020 valid tweets in total.

Methods

We started out by looking at Twitter API documentation in order to start building our query, and decided to use the tweepy Python library to collect our tweets as described in the previous section.

After all tweets were collected and shared, we began open coding of the first 200 tweets to define our topics. Since all tweets were filtered to be COVID-related, we imply any topic to be COVID-related below. For instance, *regulations* is COVID-related *regulations*. Due to the nature of our filtered tweets (see Discussion for details), we had trouble understanding the contextual meaning of some tweets. Instead of deleting them, we decided to include them since they may be significant for topic counting and TF-IDF. We will refer to them as OCT (Out of Context Tweets).

First, we went through the first 20 tweets individually to get a general ideas on the types of tweets, and came up with 5-7 topics on our own. After briefly discussing the differences and similarities in the topics that we chose, we consolidated our ideas into loosely defined groups, and went through the first 20 tweets together a second time. This time, we tried to place each tweet into one of the topics that we had chosen, but this proved easier said than done. We ended up revising our topics many times while we went through about the first 100 tweets, trying to eliminate overlap between categories and narrow down the definitions of categories that were too broad.

For instance, after a few rounds of revision, we decided to refer to the topics defined in the WHO (World Health Organization) and the Canadian government COVID-19 websites. This leads to topics such as "Current situation", "Vaccines", "Travel", Financial and economic support", and "Public health". For our purposes, we decided to divide the topic of vaccine into regulations surrounding vaccines and opinions towards vaccines, which occurred frequently in our open encoding. We also had to make careful distinctions between the topic definition for *political* and regulations since most tweets that mentioned regulations were politically related. Taking sentiment analysis into consideration, we decided to define *political* as purely targeting political figures or parties. Let S define the set of tweets that discuss anything political and are covid-related (vaccine regulations, travel restrictions, financial support), R define the tweets that discuss regulations, and F define the tweets about economy and financial support. We define P as $S \setminus \{R \cup F\}$ assuming $S = \{R \cup F \cup P\}$. We made another important decision for deciding the topic of *general pandemic*. During open-encoding, we encountered many tweets about personal experience during the pandemic (such as being bored), and found the topic of "life in the "pandemic" relevant. However, many COVID tweets also mentioned celebrities events (such as Aaron Rodgers's toe incident) that were hard to understand out of context, so we decided to expand this topic into *general pandemic* and included OCTs in this topic instead of having an "others" topic.

Sentiment analysis is a commonly studied area in data science (Feldman 2013). Overall, if a tweet shows positive support for a topic, we would encode it as positive, and vice versa. For instance, if a tweet is complaining about vaccine regulations, then we would tag it as negative for regulations. Likewise, if a tweet is discussing the effectiveness of Pfizer vaccines, we would tag it as positive for vaccine. Note that for general pandemic, since the broad topic is "life in the pandemic", if the tweet shows negative emotions (such as cursing and complaining), we would tag it as negative. For OCTs, we assigned all of their sentiment as neutral since we believed that this method has the least effect for understanding sentiment distribution. Additionally, for tweets bodies that only included quotations **and** links to other journals/websites, we decided to encode them as having neutral sentiments. This will be further discussed below.

After the process of Annotation, Data preprocession by collecting frequency-inverse document frequency(tf-idf) score. All data has to be preprocessed by (1) replace punctuation characters with a space (2) split words by spce (3) remove all emoji and hash tags (4) remove all username, etc. Finally, only alphabetical words were remained. Preprocessed data would be cleaned by removing stopwords.Here

we apply stpwords in CountVectorizer, a package in sklearn, and we additionally add task-related stopwords such as covid, pandemic, etc to extract useful information. Followed by all these procedures, we compute tf-idf score, a statistical measure that evaluates how relevant a word is to a document in a collection of documents.Following above procedures, we successfully gain the 10 words in each category with the highest tf-idf score to interpret useful information.

Results

1. Topic Definition

We divided out topics into 6 different categories based on the collected tweets: *vaccination*, *regulations*, *financial*, *political*, *COVID itself*, and *general pandemic*.

vaccination relates to a tweet's attitude towards COVID-vaccines. This includes the effectiveness and safety of the vaccine and its ability. A tweet under this category is positive if it praises the vaccine's ability, and negative if it displays vaccine-hesitancy. All "anti-vax" tweets would be interpreted as negative. For instance, the following tweet "Cardiac Dangers of the Covid Vaccine" would also be classified under *vaccination* and be tagged as negative.

regulations tweets discuss regulations related to COVID such as social-distancing, mask-mandates, and vaccine-mandates. Note here that tweets mentioning vaccine regulations are classified under *regulations*, not *vaccination*. A *regulations* is tagged as positive if we interpret it as encouraging regulations, such as vaccine mandates, and negative otherwise. All "anti-mask" tweets would be classified as negative under the *regulations* section.

financial tweets discuss the general economic situation including bonds, equity markets, money markets, etc. It also contains government economic support policy during the pandemic such as reimbursement for COVID testing, vaccine, and medical cost. This topic can reflects the economic situation during the pandemic of the society. If the tweet is discussing economic growth under COVID, then it would be classified as positive, negative otherwise. For instance, the following tweet "Almost 20% of Americans lost their entire life savings during the pandemic..." would be classified under *financial*. However, note that since it only contains quotes and links, we would tag it as neutral for sentiment analysis.

political are generally tweets that target a political figure or party without mentioning topics under regulations or financial. Note that due to the nature of twitter, most of these tweets were aggressive and tagged as negative for sentiment analysis. For instance, "@YahooNews I thought when he became President , Joe Biden would get Covid under control. Huh" would be a *political* tweet with negative sentiment.

COVID itself are tweets that talk about medical information of the COVID-19 virus or the state of the pandemic. We encountered many tweets that discuss the omnicron variant, whether about the transmission rate or number of cases in certain regions. We classified these under *COVID itself*. For sentiment analysis, we tagged a tweet in this category as positive if it expressed concerns for the virus or the pandemic, and negative if it expressed no concern. For instance, "The truth is that getting COVID didn't really help me stop overthinking COVID situations" would be a *COVID itself* tweet with positive sentiment.

general pandemic are tweets about "life in COVID", or OCT. Ranging from being bored during quarantine, reflecting on growth during the pandemic to celebrity quarantine schedules. If the tweet seems "happy", then we would tag it as positive. For instance, "My cat is SO fucking cute. Everyday she saves my life in this pandemic dumpster fire. Happy Hanukkah!" would a *general pandemic* with positive sentiment.

2. Topic Characterization

Table 1: Top 10 words by Pre-existing Topics

Ranking	Financial	General Pandemic	Political
1	debt	afraid	handle
2	stats	coat	recovery
3	exploded	strong	curing
4	shared	hat	living
5	trump	comfortable	fear
6	right	risk	stupid
7	unemployed	cold	care
8	entire	mask	democrats
9	credit	fear	planning
10	organization	peacefully	power

Table 2: Top 10 words by New Topics

Ranking	Regulations	Vaccination	Covid itself
1	order	accomplished	cough
2	governance	pfizer	taking
3	pr	admission	recover
4	preventive	begs	worry
5	discipline	evidently	tell
6	video	strategies	waiting
7	sense	hospitalisations	shadow
8	expect	question	july
9	health	jabbed	forget
10	travel	second	care

Here we characterize the top 10 frequent words appeared in each topic by their tf-idf score. Firstly, we focus on what the COVID-19 would influence people's normal life, and therefore we focused on long-exsiting social issues from Financial, Political aspect and the general pandemic aspect. The result is given in Table 1 below.

Besides, we subcategorize topics into "new changes to our life", which is problem we do not encounter before COVID times, such as regulations, vaccination, and the covid virus itself. Those topics adds new challenges to our daily life, and maybe strange for us before 2019 but become a routine now.

The top 10 words of those new topics bring by the COVID-19 is shown in Table 2.

3. Topic Engagement

In Figure 1, we show the relative number of tweets that were in each of the six categories. We see that the *general pandemic* category is the largest topic, containing 324 tweets about one-third of the total dataset. The categories of *covid itself* (221 tweets) and *vaccination* (195 tweets) are the next largest, followed by *regulations* (133 tweets), *political* (108), *financial* (39 tweets).

These results are not so surprising, since the *general pandemic* category has the broadest definition for its topic. Although *covid itself* is more specific, it is reasonable that a large quantity of tweets are discussions around the actual disease. The fact that the next largest category is *vaccination* is significant because it shows that aside from talking about the pandemic and COVID in general, the most relevant topic that many people are active in is vaccination.

The number of tweets of each sentiment per topic is visualized in Figure 2, and the percentage of the each sentiment in each topic is visualized in Figure 3. As mentioned, it is not surprising that *political* has the highest negative sentiment percentage of 62%, as many users tweet their frustrations with a political party or figure. Additionally, it contains the smallest percentage (7.4%) of positive tweets out of all six categories. *General pandemic* having high negative sentiment is not surprising either. It is interesting that *covid_itself* has a high percentage of tweets (20.8%) having positive sentiment. The *financial* category had a substantial percentage of neutral tweets (53.8%) and equal percentages of positive and negative ones (both 23.1%).

Another intriguing observation is the increased rate of non-neutral tweets in the *vaccination* and *regulations* categories. Both topics had relatively high percentages of both positive sentiment (25.6% and 31.6% respectively) and negative sentiment (29.7% and 24.8% respectively) with a lower percentage of neutral tweets compared to topics such as *general pandemic* and *covid itself*.

Combined with Tables 1 and 2, we can see the corresponding connection between the top TF-IDF words and sentiment in each topic. In general pandemic, the word ”afraid” likely resulted in the sentiment being annotated as negative. Despite having a high percentage of neutral tweets, *financial* had high TF-IDF words that would be associated with negative sentiments such as ”debt” and ”unemployed”. Similarly, *political* TF-IDF has words that would be associated with negative sentiments such as ”fear” and ”stupid”.

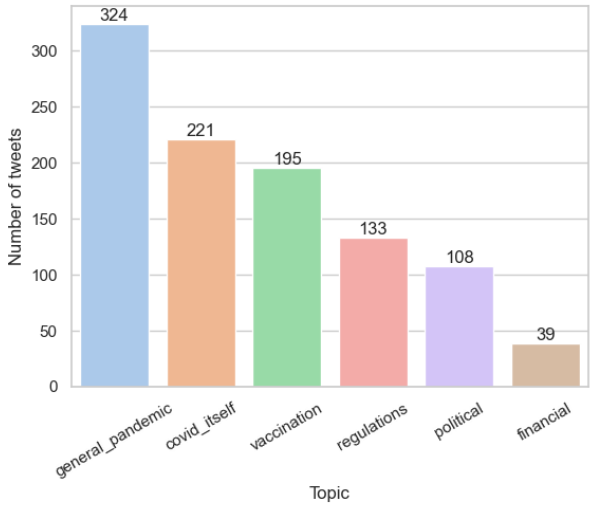

Figure 1: Number of tweets per topic.

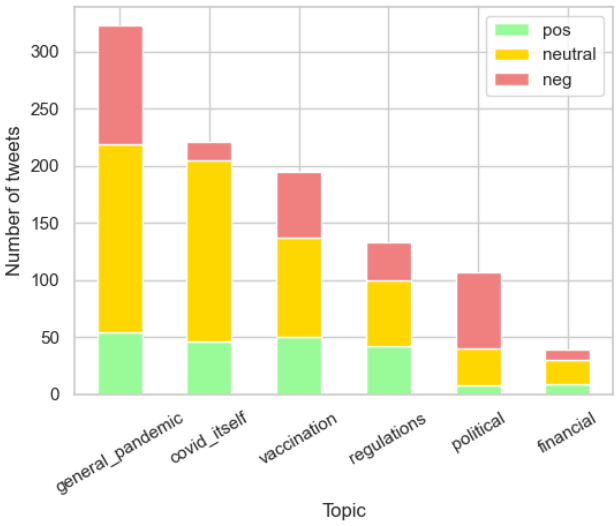

Figure 2: Number of tweets of each sentiment per topic.

Discussion

1. Result Interpretation

To further understand our results, it is important to first delve into the nature of our gathered tweets. While we only selected for uniquely IDed tweets that contained COVID-related keywords in the tweet body, we discovered duplicated tweets. This occurs when users posts the same content multiple times. An example of this is the following tweet

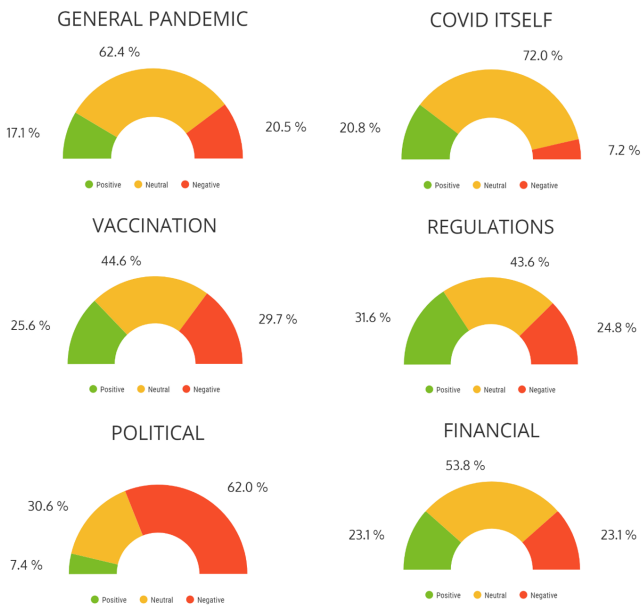

Figure 3: Percentage of tweets of each sentiment per topic.

body that was filtered into our selection 4 times with different twitter user mentions: "Another dangerous disaster is coming in the world. Your care. A terrible war is about to begin in India, China and Pakistan. People will forget coronavirus. Scientists issued alert". This can result in overcounting in topic engagement, resulting in a skewed distribution.

Another potential confounding factor in our results is out of context sentiment analysis. When we encountered tweets bodies with only quotes and hyperlinks, we annotated their sentiments as neutral. This explains why despite having high neutral sentiment percentage, *financial* contained words with high TF-IDF that would usually indicate negative-sentiments. Although it is also reasonable to infer the sentiment of a tweet based on its quotes, We decide against it since given the casual nature of twitter, tweets could be replies to users' own tweets and/or the user could be news sources. For example, an anti-vax user could first tweet quotes about the vaccine being effective, then post another tweet in reply to the previous quote, describing the information as nonsense.

The nature of our topic definition deserves further discussion. While we separated tweets that discuss the vaccine itself and its regulations, we lost the ability to distinguish vaccine-specific regulations and other regulations. We realized that both topics (vaccine itself and its regulations) can provide insight into vaccine hesitancy. During annotation, we found it difficult to distinguish some tweets from *regulations* and *vaccine*. Intuitively, a major reason behind users' vaccine-hesitancy can be vaccine safety and efficacy. While we tried our best to distinguish tweets between these 2 topics, we believe it might have been a better approach if we made a new topic for *vaccine regulations*. However, this approach would be purely driven by the goal of our project, introducing some biases. Our current approach could provide more general grouping of twitter users of pro vs. against regulations. We could also perform further analysis on *regulations* to discover the percentage of regulations tweets that are vaccine-specific.

Overall, assuming our topic engagement is an accurate reflection of twitter users' opinions, we can conclude that vaccine-hesitancy is a fairly common issue. Since 29.7% of vaccination-related tweets had negative sentiment. Additionally, based on discussion above, we performed preliminary analysis on the percentage of vaccine-specific regulations in *regulations*. We found that 44% of negative regulation tweets were vaccine-related, and 54% of positive regulation tweets were vaccine-related, which means that 17.0% of regulations tweets expressed pro-vaccine sentiment, whereas 10.9% of regulations tweets expressed anti-vaccine sentiment. Taking this into consideration, the overall vaccine sentiment might be relatively neutral.

2. Powerful Tool for Text Analysis

In this project, we explored the possibility of the methods acting as a tool for text interpretation, and showed the power of web scrapping and data analysis from the result.

It is clear that the people's altitude and thoughts can be seen from the top tf-idf score words. In regards to the *financial* discussion, top words include debt, shared, unemployed, which may indicate that many are under financial pressure and are losing jobs and that the economy is negatively impacted by the COVID-19 pandemic. For the *general pandemic* category, clothes such as coat and hat are mentioned, as well as masks to prevent the transmission of viruses. We can also see that people are very anxious about the current situation from the high frequent words afraid and fear. For the *political* aspect, it is apparent that the conversation is closely related to the recovery rate, and people are concerned their livelihoods. However, they may generally have a negative altitude towards the government or the policy, which result in words like "stupid" and "fear" and correspond to the high negative sentiment rate(62%) for tweets under the political topic. This further proves the potential of this methods to be a useful tool for text analysis. Looking at the *regulations* topic, we can see the travel restrictions have a large impact on people's lives, and this might be why 'pr', the abbreviation of permanent residents, receives a high score. The regulations has changed people's lifestyle with videos instead real meeting, and people are expecting back to normal. Here we can see most words does not have emotional attitudes,corresponding with around half of tweets in this topic are neutral(43%). For the vaccination, we can see what people are concerning- the brand of vaccination,the increasing hospitality rate, and the second shoot. Also, people have mixed feeling about covid itself. Integrating all these information, we can get many interpretable information and hence we conclude that it consistent with the sentiment analysis, and showed the power of being an text analysis tool.

3. Limitations

However, this method also has drawbacks, and there is a large space for future improvement. Since the methodology is split into three parts: data collection, data annotation and data analysis, the flaws come from these two aspects.

Firstly, the process of open-coding is done by one group member due to the limitation of time and resources. However, the result would be more reliable with multiple members coding on the same task and getting a similar effect, validating the correctness of the code. Also, we filtered out those tweets containing the keywords, but with keywords in links directly; however, we can have some exciting findings if we go deeper into those links. Quoting other links or websites is normal in reality, and discarding them contradicts what we behave in normality and may slightly influence the result. Lastly, many tweets appeared more than once when collecting data, and keeping them or not remains a tricky problem for the trade-off between redundant analysis and closeness to real-life settings.

Secondly, the process of Annotation is done manually by Mark Bai. Manual Annotation is an objective task, and since analysis relies on annotation data, it may largely influence the reliability and interpretability of data. It is easy for one tweet to fall in the intersection of 'positive/negative' and 'middle,' and there is also a "catch-all" category(i.e., general pandemic). Since it is manually done, there is no such threshold or criterion for classification, and also impossible for method reproduction. Despite its unreliability and instability, it is also very time-consuming.

Last but not least, the analysis of data was done by python program automatically, but we can set stopwords and filter out unwanted words based on the needs. For example, in this task, terms like "pandemic" and "covid" are set as stopwords since they contain keywords in the topic, and therefore the tf-IDF would be undoubtedly high but cannot give us an interpretable meaning. Meanwhile, this also causes the methods to be too flexible and objective, and there is a possibility that the interpretation of texts through data analysis is distorted and overobjective.

4. Future Improvements

Given the above limitations, the improvement could be made to improve the performance and reasonability of this model. Firstly, the manual task should be done by multiple members and then take the average or voting to ensure the subjectivity of data and Annotation. Secondly, natural language processing techniques(NLP) should be considered to perform automatic Annotation with clear standards and thus guarantee reproduction. Benefitting from the high-throughout annotation ability, the improved method would apply to large datasets with higher efficiency.

Group Member Contributions

In this project, every group member contributed equally. Sophia did the script writing task to collect tweets and compiled them in the shared spreadsheet for the group to work on. After doing the open-coding part, Mark was responsible for the annotation of topic and sentiment with positive, negative and neutral. Sophia also assisted in data annotation. Nicole Xu was in charge of computing tf-idf score and finding the top 10 frequent words. Everyone worked together to define the topic definitions and typology for each topic and contributed equally to the report writing.

References

Feldman, R. 2013. Techniques and applications for sentiment analysis. *Communications of the ACM* 56(4):82–89.

Kraemer, M. U.; Yang, C.-H.; Gutierrez, B.; Wu, C.-H.; Klein, B.; Pigott, D. M.; Group†, O. C.-. D. W.; du Plessis, L.; Faria, N. R.; Li, R.; et al. 2020. The effect of human mobility and control measures on the covid-19 epidemic in china. *Science* 368(6490):493–497.

Public Health Agency of Canada. 2021. Covid-19 vaccine doses administered in canada. `https://health-infobase.canada.ca/covid-19/vaccine-administration/`.