CSC110 Project Report: Analyzing Vaccination Rates and Related Tweets

Allen Uy, Nathaniel Yong, William Yao

Tuesday, December 14, 2021

Problem Description and Research Question

Public administration of COVID-19 vaccines in Canada first began en masse in December 2020 (Aiello & Forani, 2020). In the year since then, over 80% of Canada's population over the age of 12 have become fully vaccinated against the virus (Dunham, 2021)—marking it as Canada's most ambitious and impactful immunization project in history.

Although logistical issues plagued Canada's early vaccine rollout, the central obstacle to increasing national vaccination coverage has since shifted to vaccine hesitancy among many Canadians who remain unvaccinated. According to the Canadian Community Health Survey conducted in late 2020, 23.1% of Canadians over the age of 12 indicated an unwillingness to take the vaccine (Government of Canada, 2021). Hesitancy rates have declined since then, but they remain a significant barrier to achieving the vaccination threshold necessary for nationwide herd immunity and reopening activities with full confidence.

As with all recent social issues, much of the civil discourse surrounding COVID-19 has taken place online via Internet forums and social media platforms. This is especially true of the pandemic because social distancing has prompted increased online interaction for everyone, and also rendered many other channels of public communication infeasible. Popular social networks like Twitter are increasingly assuming the role of an online "town square" containing discussions that accurately reflect general public sentiments. For this reason, governments and scientific bodies employ social media to not just disseminate vital information on public health crises, but also analyze patterns in civil discourse that foreshadow crisis developments and public reaction to new policies. Many political movements, including both pro-vaccine and anti-vaccine groups, also owe the spread of their ideology to social media.

The establishment of these online "town squares" entails a technically potent side effect; because public communication is now digitized, computer scientists can collect these communications as a data set for studying public attitudes at unprecedented scale and efficiency. One effective tool for this is analyzing the online frequency of keywords associated with anti-vaccine sentiments. For this, we have selected Twitter as the representative online platform from which we will extract historical user-generated posts and measure keywords. We therefore define our research question as follows: How has the frequency of keywords associated with anti-vaccine sentiment evolved over time? Can this frequency create a useful predictive model for vaccination in a population? If so, we hope that health policymakers can use trends identified in our research to better predict the coverage outcome of a given vaccination strategy, and focus public information campaign resources on regions that are most susceptible to vaccine hesitancy.

Dataset Description

We used three datasets for this project. The first one contains vaccination data provided by the Centers for Disease Control and Prevention (CDC), the second one contains vaccination data provided by the Government of Canada, and the third contains tweet-ids related to anti-vaccine content, provided by Muric et al. The first two are .csv files while the tweet-ids are contained in multiple text files organized by date.

The CDC dataset encompasses all necessary information on vaccination statistics in the United States. It is extremely detailed, containing 69 columns that store information (by both date and individual state) on the number and type of vaccines administered, as well as other information like per capita doses and doses by age range. The detail found in this data set allows for precise analysis of vaccination coverage in the U.S., but most of it was not used for this project given the sheer breadth of the data set.

The Government of Canada dataset encompasses all necessary information on vaccination statistics in Canada. It contains fewer columns and more generalized data, being: date, vaccines administered, number by dosage, and delta change. While this means that less precise analysis of vaccination coverage is possible, the information is sufficiently detailed for the scope of our project.

The tweets were gathered by monitoring a list of keywords related to anti-vaccine ideas. The dataset was filtered to specifically only include English tweets, with the majority being posted by Internet users in North America. In order to comply with Twitter guidelines, the provided dataset contains only the tweet-ids of tweets collected from 2020-10 to 2021-11. A representative sample of approximately 30,000 tweets were used in the project, ranging from 2020-10 to 2021-04 for the sake of fast computation.

Computational Plan

Our program reads csv files and json files to aggregate and transform data about COVID-19 vaccinations statistics as well as tweets related to vaccinations.

In vaccination_data_loader.py, csv files related to vaccination data are converted into pandas dataframes with the create_df_from_csv function. The functions convert_df_types_canada and convert_df_types_cdc also helps to clean the data by converting the rows to more appropriate data types. For example, using several helper functions, it converts the date strings to python datetime objects. Also, because we are collecting vaccination data from both the Government of Canada and the CDC, we use the sum_dfs function to aggregate both dataframes into one dataframe.

In tweet_data_loader.py, jsonl files with tweet data from Twitter are converted to pandas dataframes. The read_files function is used to iterate over all the jsonl files in the specified directories, and aggregate all the data into one list. These jsonl files contain the relevant json for each tweet fetched by the Twitter API. They are opened with gzip since they can be rather large. This function uses the read_jsonl and process_json functions to convert the jsonl format into a list of python dictionaries containing the tweet data. Finally, the create_df function converts the list of dictionaries into a pandas dataframe while also helping to clean the data by converting certain rows to more appropriate data types. Using the convert_str_to_datetime, the date strings are converted to python datetime objects using the specific formatting.

In tweet_computation.py, we filtered tweet data to obtain more relevant information. The remove_stopwords function takes in a string and removes all stopwords, which are words that don't contribute to the meaning of the text. The create_words_and_hashtag_df takes the dataframe containing the tweet data from tweet_data_loader.py and filters it into two dataframes, the first containing the most commonly found words in the tweet data and the second dataframe containing the most common hashtags, using the Counter object from collections. The remove_stopwords function is also used here to ensure that the less relevant words are discarded.

Finally, in main.py we visualize the data. Two bar graphs are created using the dataframes from tweet_computation.py. These bar graphs show how many times the most common words and hashtags occurred in the tweet_data. Using the vaccination dataframes from vaccination_data_loader.py, a line graph is created which shows the total vaccinations in both Canada and the US over time.

Instructions about how to obtain datasets and run program

To obtain the CDC vaccination data, go to https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/unsk-b7fc. Click on “Export” in the top right, and then download the CSV file.

To obtain the Government of Canada vaccination data, go to https://health-infobase.canada.ca/covid-19/vaccine-administration/. Scroll down to where it says “Doses Administered Nationally” and then click on the “Access the data” button next to it to download the CSV file.

To obtain the raw tweet-id dataset, the simplest way would be going to https://github.com/gmuric/avax-tweets-dataset and cloning the repo. IMPORTANT: the original dataset should not be downloaded. We have provided the processed dataset in UTSend along which contains the necessary zipped json files and the text files containing the tweet-ids which were provided by the original dataset, we will describe the process of using the Twitter API but will not provide other info since we should not disclose our API key or secret.

Please use the following UTSend info:
Claim ID: nZcwHgu8YobpC96T
Claim Passcode: NzYTiKWG35XGGBBV

To ensure the proper directory levels, all python files should be at the top level. The vaccination data provided by the CDC and Canada should be placed under data/vaccination. The tweet dataset should be unzipped under data/avax. After unzipping you should see data/avax/streaming-tweetids with 7 subfolders corresponding to date. Each folder contains many text files and zipped files, nothing further has to be done.

For completeness, a file named hydrate.py was included. This file is included in the GitHub repository of the tweet dataset and is left in the state it can be downloaded in. If you were to run this program from absolute scratch, you would register a developer account on Twitter, followed by installing twarc. Twarc is not included in our requirements.txt since we do not expect you to run it. You would setup twarc in the CLI using "twarc configure" to enter your API key and secret, cding into the proper folder "data/avax", and finally running "python hydrate.py -streaming" in the terminal. To reiterate, this file was not coded by us.

To run our program, simply run main.py. After a "short" wait (upwards of a few minutes), 3 separate tabs should open with a plotly graph. The first is a bar graph of the top 20 words in tweets, the second is a bar graph of the top 15 hashtags, the third is a line graph of total vaccinations.

Changes between proposal and final submission

Our final approach towards analyzing vaccine sentiments differed considerably from the proposed strategy in two primary respects. The first is that instead of analyzing the frequency of keywords associated with anti-vaccine sentiments, we had originally wanted to conduct sentiment analysis across all vaccine-related tweets to measure all opinions, including positive ones, towards the COVID-19 vaccine.

This original approach was more computationally ambitious, but in discussion with TAs, our team decided that it would have sacrificed too much precision for breadth. In particular, Internet users are able to express complex sentiments towards vaccines in subtle ways that cannot be accurately discerned through sentiment analysis that is simply positive or negative. Instead, we decided that analyzing specific keywords that are known to be associated with anti-vaccine sentiments would be more directly useful for identifying the issue of anti-vaccine movements facing public health policy.

Secondly, our proposed strategy included an analysis of whether vaccination rates in a particular geographic area correlated with vaccine sentiments in online posts from Twitter users residing there. This would have been an interesting relationship to explore as it would more precisely determine the relationship between these two trends, but after taking a sample of 130,000 tweets in our data set, only 950 had the relevant geolocation data provided by Twitter in the interest of user privacy. We felt that because such a small proportion of our data set had the relevant information, we would not be able to discern representative trends for different geographic locations, especially when considering areas with fewer users.

Discussion

Our final project is intended to extract and display relevant information on anti-vaccine sentiments and vaccination rates to assist public health policymakers in detecting obstacles to vaccine coverage. Running our main.py file generates 3 separate graphs that depict the information we extracted from the selected data sets. In order, these are:

1. A bar graph showing the frequency of hashtags known to be used by the anti-vaccine movement
2. A bar graph showing the frequency of keywords associated with anti-vaccine sentiments
3. A line graph showing the total doses of COVID-19 vaccine administered in Canada and the U.S. combined from 2020-12 to 2021-12

Before interpreting the results, it is important to recall that for the purpose of delivering a computationally feasible product, the tweets analyzed for the Graphs 1 and 2 are a large sample of our tweet dataset and do not encompass the entire dataset. This has two especially important implications for our discussion.

First, the 30,000 tweets from which this information was extracted range in time from 2020-10 to 2021-04, while Graph 3 depicts vaccination data from 2020-12 to 2021-12. We are particularly concerned with the intersection of these periods to try and find relationships between vaccine sentiment and actual vaccine coverage, but the incongruity in the timespans of our datasets should be considered when visually comparing the graphs. Second, because keywords were only counted for a sample of 30,000 tweets, the numbers do not portray the absolute prevalence of those anti-vaccine keywords, but rather their prevalence relative to each other.

Our team considers Graph 1 to display the most interesting information extracted by our project. In particular, we note that Twitter users with anti-vaccine sentiments made use of two specific hashtags, #FireFauci and #WeWillNotComply, far more than any others. Other major hashtags used include general statements like #NoVaccine, #NoVaccineForMe, #MyBodyMyChoice, and so on, as well as targeted statements like #BillGatesBioTerrorist and #ArrestBillGates.

We are most interested in the prominence of negative hashtags directed at specific persons of interest. These include notable figures Dr. Anthony Fauci, who has been at the public forefront for the U.S.'s COVID-19 response for virtually the entire pandemic, and Bill Gates, who has also pushed for increased vaccine coverage and is associated with common anti-vaccine conspiracies like their inclusion of "microchips" (Goodman & Carmichael, 2020).

This trend suggests that the most popular anti-vaccine hashtags are often those that attack specific figureheads with a strong public presence, as opposed to more abstract concepts like the institutions or ideas they represent. For instance, consider that the most popular hashtag is #FireFauci and not similar statements like #EndTheCDC, #EndTheNIH, or any other institutions represented by and associated with Dr. Fauci. In the context of developing public health policies for increasing vaccine coverage, this finding may indicate that the charisma and political approval of authority figures are important to encouraging acceptance.

Besides the findings of such a strong focus on individual persons of interest by the anti-vaccine movement, we also hope Graph 1 will be generally useful for summarizing the most common fears and personal principles associated with vaccine hesitancy. In particular, we note a strong emphasis on the principles of consent and personal liberty (consider #MyBodyMyChoice, #InformedConsent, #Freedom, and, among our personal favourites, one tweet with #mydogmychoice), that are often at odds with the concept of vaccine mandates. As such, reconciling these conflicts of principles will be an essential challenge for public health to tackle as further vaccine policies are implemented.

In Graph 2, we see the high frequency of the word RT, indicating a retweet. A lot of these messages are posted by one person and then spread around. Even with removing stopwords, due to the nature of tweets that typically are not fully thought out, many of the most frequent words are a bit gibberish. "I" is included since it technically is not a stopword. "(siren emoji) NEW" is a part of a sentence in a tweet that was retweeted quite a few times. The full start of it is, "(siren emoji)NEW BILL ALERT!(siren emoji) The "#FireFauci Act" slash salary Dr. Always Wrong $0 "#WeWillNotComply Act". Without a naive approach of simply splitting the string and iterating over the words, perhaps more important ones could be found. Given more time, we could implement a vector space model which can highlight the importance of words.

Finally, we can integrate our findings from Graphs 1 and 2 with Graph 3. Perhaps most interesting is that the hashtags and keywords were extracted from tweets posted between 2020-10 and 2021-04. For the majority of this time period, we see in Graph 3 that vaccine doses in Canada and the U.S. combined were rolled out very gradually without elapsing enough time to consider long-term effects. This could indicate that the aforementioned principles of the anti-vaccine movement are not based on experimental data from the public rollout, but either the experimental tests conducted by vaccine manufacturers during research and development, or preconceived notions of the COVID-19 vaccine.

Computationally speaking, our team is satisfied with the results of the project. We have gained considerable experience in manipulating new data types like the pandas dataframes, and in making connections between these objects and their abstract counterparts, given that we have explored multiple ways to implement tabular data in CSC110. We have also developed our skills in using algorithms to process raw data sets and efficiently extract a diverse range of information.

However, we are also aware of several lessons in programming to consider when further exploring this topic in the future. Most importantly, we are still excited to use geolocation as a means to more precisely determine the relationship between vaccine sentiments and actual vaccine coverage. Thus, we look forward to finding other data sets that extend beyond anti-vaccine tweets, as well as designing more computationally efficient algorithms for this increased breadth, to make location-based analyses and choropleth charts of vaccine sentiments.

References

Aiello, R., Forani, J. (2020, December 14). 'V-day': First covid-19 vaccines administered in Canada. Coronavirus. Retrieved November 6, 2021, from https://www.ctvnews.ca/health/coronavirus/v-day-first-covid-19-vaccines-administered-in-canada-1.5230184.

Al Jazeera. (2020, November 27). Most Canadians will get COVID-19 vaccine by September: Trudeau. Al Jazeera. Retrieved November 6, 2021, from https://www.aljazeera.com/news/2020/11/27/most-canadians-will-get-covid-19-vaccine-by-september-trudeau.

Dunham, J. (2021, September 28). More than 80 per cent of eligible Canadians fully vaccinated against COVID-19. Coronavirus. Retrieved December 13, 2021, from https://www.ctvnews.ca/health/coronavirus/more-than-80-per-cent-of-eligible-canadians-fully-vaccinated-against-covid-19-1.5603485.

Goodman, J., Carmichael, F. (2020, May 30). Coronavirus: Bill Gates 'microchip' conspiracy theory and other vaccine claims fact-checked. BBC News. Retrieved December 14, 2021, from https://www.bbc.com/news/52847648.

Government of Canada. (2021, March 26). COVID-19 vaccine willingness among Canadian population groups. Statistics Canada. Retrieved November 6, 2021, from https://www150.statcan.gc.ca/n1/pub/45-28-0001/2021001/article/00011-eng.htm.

Muric, G., Wu, Y., Ferrara, E. (2021). Covid-19 vaccine hesitancy on social media: Building a public Twitter data set of antivaccine content, vaccine misinformation, and conspiracies. JMIR Public Health and Surveillance. Retrieved December 14, 2021, from https://publichealth.jmir.org/2021/11/e30642.

Tsao, S. F., Chen, H., Tisseverasinghe, T., Yang, Y., Li, L., Butt, Z. A. (2021). What social media told us in the time of COVID-19: a scoping review. The Lancet. Digital health, 3(3), e175–e194. https://doi.org/10.1016/S2589-7500(20)30315-0

Centers for Disease Control and Prevention. (2021). Covid-19 vaccinations in the United States, Jurisdiction [Data File]. Retrieved November 6, 2021, from https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/unsk-b7fc.

Public Health Agency of Canada. (2021). Updates: COVID-19 vaccine doses administered in Canada [Data File]. Canada.ca. Retrieved November 6, 2021, from https://health-infobase.canada.ca/covid-19/vaccine-administration/.

Plotly Documentation. https://plotly.com/python/
Pandas Documentation. https://pandas.pydata.org/
Nltk Documentation. https://www.nltk.org/