COMP 598 Final Project: COVID in Canada

Emily Tam[1], **Veronica Tarka**[1], **Sonia Boscenco**[1]

[1]McGill University
emily.yw.tam@mail.mcgill.ca, veronica.tarka@mail.mcgill.ca, sonia.boscenco@mail.mcgill.ca

Introduction

The COVID-19 pandemic has caused over five million deaths globally and dramatically interrupted individuals' lives in matters ranging from education to business to childcare [7]. Since the World Health Organization approved a vaccination against COVID-19 for emergency use in January 2021, a global effort to vaccinate the world's population has been underway [2]. This vaccination campaign affects every individual in the world and has thus resulted in a contentious conversation about the safety, ethics, and politics surrounding the vaccine and vaccination efforts.

We aimed to capture this conversation through an analysis of 1,000 tweets pertaining to the pandemic and the vaccine against COVID-19. We sorted the tweets into categories distinguished by purpose (e.g. sharing a fact versus an opinion) and setting (e.g. talking about work versus the government), and we further subdivded these sections based on the sentiment contained within: positive, negative, or neutral.

We found that several categories presented interesting contrasts that may serve as windows to the different spheres of life. The category designed to capture factual reports (the 'info' category) on COVID-19 contrasted strongly against the category for opinions about COVID-19. Factual reporting, while not immune to negative sentiment, showed significantly less negative sentiment than opinions (21% versus 70%). Opinionated content also showed significantly less positive sentiment than tweets pertaining to the Twitter user's personal life (8% versus 28%). These two observations may reflect a dramatic bias towards sharing negative opinions about COVID-19 which are not necessarily founded in users' personal lives. The 'info' and 'opinions' categories made up 30% and 25% of all the tweets, respectively.

We found that the workplace does not appear to be a particularly common setting to prompt discussions about the COVID-19 or COVID-19 vaccines as it comprised only 2% of the sample. Similarly, content related to schools and children only comprised 3%. Conversations surrounding politics, especially Donald Trump, vaccine mandates, and border closures prompted much more discussion than schools and work, comprising 18% of the dataset, but content centered around politics did not exhibit the same dramatic bias towards negative sentiment as opinions not directly related to politicians or governments.

Our analysis identified key areas of Twitter content that contained emotional language about COVID-19 and the vaccine, and uncovered a bias towards negative opinion-sharing. Future work should consider using a larger dataset collected over more time to determine whether these results hold more generally.

Data

We collected 1,210 tweets from Twitter over the course of the three days falling between December 5 and 7, 2021, using the content request protocols supported by the Twitter API v2. We searched the most recent tweets published immediately before the time of collection without regard to author or location. All tweets collected were tagged as English ('en') content by Twitter.

Each of the collected tweets contained one or more of the following key words either in the main content, as a hashtag, or in the title of a linked article: COVID-19, vaccination, Pfizer, Moderna, Astrazeneca, or vaccine. The keywords were matched without regard to letter case. Other brands of vaccines against COVID-19, such as Sputnik or Johnson & Johnson, were not included due to the fact that they also denote many non-vaccine-related topics like satellites and baby powder. We did not include 'coronavirus' as a keyword because it narrowly describes the virus causing the COVID-19 disease, and our analysis is focused on the vaccine against the COVID-19 disease. Slang terms for the vaccine like 'shot' and 'jab' were not included as keywords as they also describe acts of violence unrelated to our focus. We also excluded the slang term 'vax' (short for 'vaccine') as a keyword because it is most often found in the phrase 'anti-vax' describing anti-vaccination views, and we aimed to collect a set of tweets unbiased towards a particular stance on COVID vaccinations. Additionally, 'anti-vax' includes anti-vaccination views associated with other vaccines besides COVID-19 vaccines.

Retweets and tweets posted as replies to other tweets were filtered from the dataset to eliminate repeated content. This largely eliminated identical content, but some duplicate tweets remained if they were not tagged as retweets or replies by Twitter. This is common when news organi-

zations post from their many affiliate accounts (e.g. BBC News, BBC World, BBC Sport) about the same news story using only the article's title as the content. To address this, we manually checked the collected tweets for identical content posted close together in time and eliminated all of the duplicate tweets.

Next, we cleaned the dataset in order to filter out tweets that had made it through our initial filter. The 'en' tag did not filter out all tweets from other languages so we marked them as 'discard' during annotation and dropped them from the csv file. We also dropped some tweets in English that were uninterpretable (e.g. "Doing Trump dirty work? What Covid didn't accomplish they will try"), which might have been written by bots and did not fit into any of the categories we had established. We had also used a few different annotations for 'positive' sentiment, as it was not possible to use our initial coding of '+' in xlsx files so we standardized all of the sentiment annotations to be 'positive', 'neutral', and 'negative'.

Finally, we tokenized all the tweets using the TweetTokenizer() package from nltk. We found that this was the most accurate in terms of tokenizing words, hashtags, links, and users correctly. We had initially tried to manually tokenize the tweets by lowercasing the words, replacing punctuation with spaces, and splitting by spaces as we had done in previous assignments. However, this did not properly tokenize the tweets and led to many mistakes in the final tokenized tweets. After tokenizing, we obtained a list of words that occurred at least two times across all tweets. We then filtered stopwords, using the stopwords file provided in Assignment 8. We removed the '#' symbol from hashtags so that words matching a hashtag would be counted together. We also filtered out any words containing non-alphabetic characters (anything that is not 'a-z'). After collecting the word list, we manually went through the list to ensure that it made sense and properly filtered everything out. We found some hashtags that had made it through the filtering process that we then removed, along with certain contractions that had not been included in the stopwords file and non-English words that had been missed previously. The total number of cleaned and filtered tweets was 1163 and there were 1807 words in our final word list.

Methods

We manually annotated the collected tweets based on six main categories: politics, school, workplace, public health information, personal life, opinions. The categories were formed using the first 200 tweets. Tweets that contained any mention of a political figure or governing body like a parliament, from any country, were annotated as 'politics'. Content mentioning government intervention or policies like border closure and vaccine mandates were also included in this category. Tweets in the workplace category fell under five main categories: (1) they mentioned a COVID-19 outbreak at their work, (2) they discussed personal opinions on vaccine mandates at their work, (3) they mentioned working at their home office, (4) they mentioned the impact of COVID-19 on their business, (5) they mentioned the impact of COVID-19 on the stock market. Tweets pertaining to staff or students of elementary schools, high schools or higher education were included in the 'school' category. The 'personal life' category was comprised of any tweets that described an event in the author's personal life that was related to COVID. This included getting COVID-19, a family member that had COVID-19, going to music concerts where COVID-19 spreading is a concern, etc. The 'public health information' and 'opinions' categories occupied distinct spheres of Twitter content. Although they both contained tweets that aimed to inform the public about COVID-19 and related topics, the defining difference between the two categories was whether or not the information was primarily delivering facts or was a single person's opinion. For example, any COVID-19 statistics or information regarding COVID-19 or COVID-19 vaccines that came from peer-reviewed articles, public health units, or credible news sources (e.g. New York Times, Globe and Mail, CBC) was included in the 'public health information' category. Any tweet that was a single person's opinion or recommendation (with the exception of a recommendation coming from Dr. Anthony Fauci or other public health experts, whose recommendations are based in scientific fact), was categorized as a personal recommendation. Additionally, any news that was not from a reputable source (e.g. conservatives.org, Fox News) and was not based on peer-reviewed articles was included in this category, as were any reactions to news articles or recent COVID-19 mandates. Tweet links were followed to verify the reputability of the source.

If a tweet could have been annotated as part of two or more categories, it was assigned to the category that it fit the best. For example, the tweet 'Getting ready for my holiday work party - hope there's no COVID-19 spreading' could fit either in the workplace or in the personal life categories, but the general idea resolves around the workplace. Posts were also annotated for sentiment: either positive, negative or neutral. A post was considered neutral if it was a fact (e.g. COVID-19 case counts) or did not contain any words denoting feeling or emotion. Negative sentiments included anger, frustration, and sadness. Positive sentiments included humour and overall happiness or joy (e.g. 'I'm happy I have finally recovered from COVID!').

To compute tf-idf, we started with a word list containing words that occurred at least 2 times across all cleaned and filtered tweets. We then obtained word counts for each topic, and the total number of words from tweets from each topic. The tf-idf values for each word for each topic was calculated using the following formulas:

$$\text{tf(t, d)} = (\text{\# of times t occurs in d})/(\text{total \# of t in d})$$

$$\text{idf(t, d)} = log((\text{total \# of d})/(\text{\# of d where t appears}))$$

$$\text{tfidf} = \text{tf} \times \text{idf}$$

In these formulas, each term (t) is a word from the word list and each document (d) is one topic. Once TF-DF values were calculated, we sorted them to obtain the top 10 words in each topic category.

Results

The results of the tf-idf analysis provided words for each topic that were appropriate given the topic definitions (re-

Topic	Definition	Top 10 words by tf-idf score
Info	• COVID-19 statistics • information regarding COVID or vaccines that came from peer-reviewed articles, public health units, or credible news sources (e.g. New York Times, Globe and Mail, CBC, etc.)	reports, uk, insights, analytics, pericarditis, omicron, rise, infection, player, ben
Opinions	• single person's opinion or recommendation (except for recommendations coming from Dr. Anthony Fauci or other public health experts) • news providing misinformation (not based on peer-reviewed articles/scientific facts) • reactions to news articles or recent COVID mandates	govts, abuse, people, costly, advisors, twitter, killed, naturally, world, flu
Personal	• describe an event in user's personal life that relates to COVID (e.g. getting COVID, family member getting COVID, concerns about spreading COVID at a concert, etc.)	kody, little, lol, mom, smell, hotel, booster, fever, bed, trash
Politics	• any mention of a political figure or governing body (e.g. parliament), from any country • government intervention or policies (e.g. border closure, vaccine mandates) • COVID-related protests	blasio, mayor, sector, mandate, employers, queensland, city, reopen, trump, nyc
School	• pertaining to staff or students of elementary schools, high schools, or higher education	school, students, schools, exams, class, institutions, return, dressed, quality, diseases
Workplace	• any mention of a COVID outbreak at work • opinions on vaccine mandates at work • working at a home office during COVID • impact of COVID on business • impact of COVID on the stock market and economy	aviation, profitable, omg, investors, customers, blessed, company, businesses, workers, air

Figure 1: Topic definitions and the top 10 words by tf-idf score for each topic

.

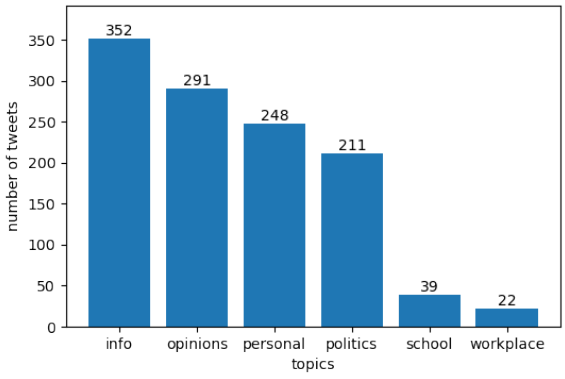

Figure 2: Number of tweets in each category using all tweets

.

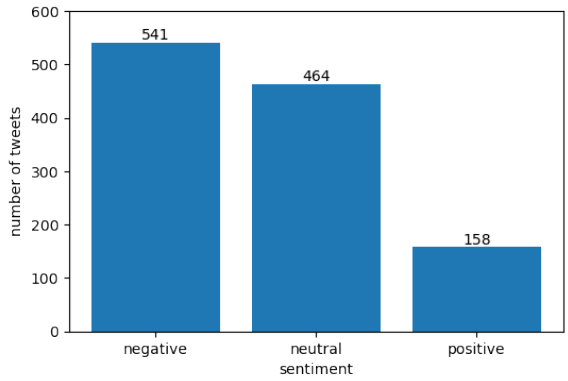

Figure 3: Number of tweets by sentiment using all tweets

.

fer to Figure 1). We also graphed the number of tweets in each category (Figure 2); the number of negative, neutral, and positive tweets (Figure 3); and the number of positive, negative, and neutral tweets in each category (Figure 4).

In order to understand discussion happening around COVID-19 vaccine hesitancy, we also analyzed a subset of the tweets containing vaccine-related words, such as 'vaccine', 'vaccination', 'pfizer', 'moderna', etc. found in the tweets. There were 509 vaccine-related tweets in total. After conducting the same tf-idf analysis on this subset, we obtained very similar results that were not very helpful in assessing vaccine hesitancy. However, we also conducted a sentiment analysis of the vaccine-related tweets, and found that most of the tweets in this subset were negative (235 out of 509), rather than neutral (204 out of 509) or positive (70 out of 509), suggesting that vaccine hesitancy was likely high in the tweets we collected (refer to Figure 5). It was also interesting to find that the majority of vaccine-related tweets in the 'opinions' category (110 out of 104) were negative (refer to Figure 6).

Discussion

Public Health Information

This category showed the highest engagement, making up 30% of the total tweets. This is expected, as Twitter is a hub for news traffic. The two categories with the most tweets, public health information (352 tweets) and opinions (291 tweets) displayed stark differences in sentiment. We noticed that the majority of 'info' tweets are of neutral sentiment (68%); this is not surprising as the category was built to capture factual content. This is also supported by the top ten words containing 'reports', 'insights', 'analytics', 'uk', 'rise', and 'omicron'. These words depict the rise in reporting on the Omicron coronavirus variant currently unsettling communities across the world. Moreover, the UK is overrepresented in this category due to the influx of Omicron-related tweets in response to the recent rise in COVID-19 cases and restrictions in the UK [6]. The 'rise' and 'infection' keyword suggest that tweets containing or about science-based articles, along with COVID statistics, are primarily concerned with whether COVID infection cases are

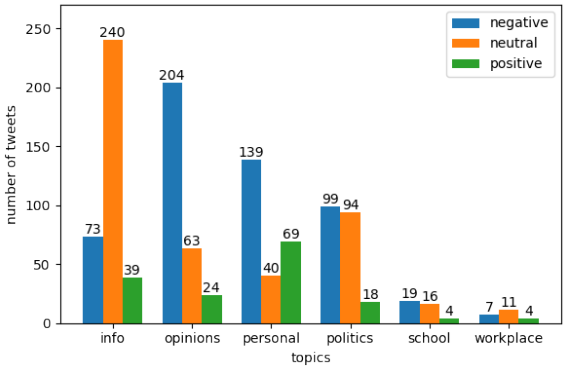

Figure 4: Number of negative, neutral, and positive tweets in each category using all tweets

.

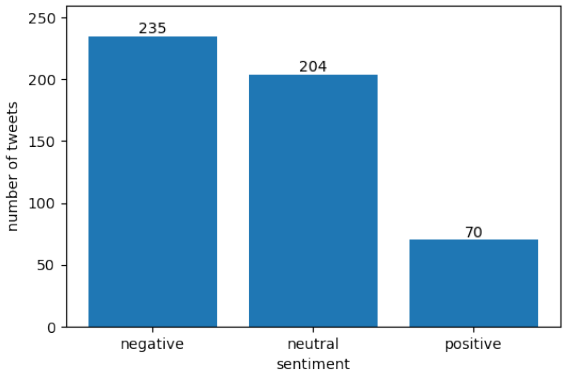

Figure 5: Number of tweets by sentiment using only vaccine-related tweets

.

rising and what the impact of this is. We also noticed that 'ben', 'player', and 'pericarditis' were among the top ten words. It is not initially obvious why words were included in the top 10 words by tf-idf value. However, after searching through tweets containing these keywords, we found 6 tweets including news articles about how the Australian basketball player Ben Magden was diagnosed with pericarditis after taking his second shot of the Pfizer vaccine. These words likely appeared in the top words due to their uniqueness in the topics, as tf-idf values increase proportionally to the number of times a word appears in a document and is offset by the number of documents that contain the word.

Although this category was built to capture factual reporting, 21% of the content was negative and 39% was positive, indicating emotion-ridden and possibly biased reporting. Many tweets were statistical reports of COVID-19 cases or vaccination rates, which made up a large proportion of the neutral sentiment tweets. An example of factual yet emotional content is, 'A shocking leap in cases of COVID-19 devastates Kentucky's labour market.' This is a fact, not an opinion, but keywords like 'shocking' and 'devastates' indicate a negative sentiment. Furthermore, the Ben Magden

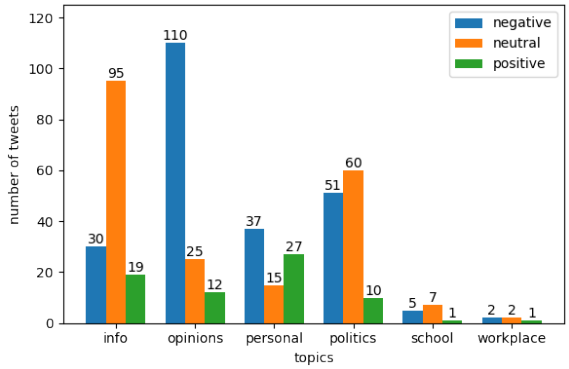

Figure 6: Number of negative, neutral, and positive tweets in each category using only vaccine-related tweets

.

tweets had negative sentiments as they expressed shock or concern in response to his vaccination-induced condition.

Personal Opinions and Recommendations

In the opinions category, negative sentiment (70%) tweets greatly outnumbered both neutral (22%) and positive sentiment (8%) tweets. This category also had the second-highest engagement, accounting for 25% of all tweets (291 total). This high engagement, complemented by the large proportion of negative sentiment tweets could be attributed to a greater number of people using Twitter as a way to voice their anger or concerns about COVID-19 related issues compared to those who simply want to share their opinions. We observe that 'abuse', 'costly', and 'killed' are among the top ten words. These are strong words that themselves hold negative sentiments. One tweet mentioning 'abuse' reads "I've said it before and I'll say it again: these covid restrictions are abuse. Telling people what to wear is abuse. Forcing them to get a "vaccine" from a corrupt company is abuse. All of this garbage is abuse." From this example, we can gather that tweets in the 'opinions' category focused on individual people's issues with COVID restrictions, mask restrictions, and vaccine requirements. We also notice the words 'naturally' and 'flu' appear frequently in this category. Many tweets compared COVID-19 to the flu and were accounts of people that felt as though their immune system could naturally fight off any virus, including the COVID-19. Much of the content contained anger over public health decisions, intense suspicion of the government, and comments intended to place blame. This is reflected by the overwhelmingly negative sentiment found in this category's tweets and by keywords 'govts' and 'world' coupled with the abovementioned 'abuse' and 'costly' indicating a general mistrust of the world's leaders.

It is important to contrast the personal opinions category with the public health information category. Many Twitter users struggle to distinguish factual information from personal opinions. This analysis shows that the content of these two spheres is fundamentally different in both sentiment and content, and should be considered differently in the readers'

minds to avoid continued spread of misinformation through opinions masquerading as facts about the COVID-19 and the vaccine against COVID-19.

Politics

This category accounted for 18% of the total tweets. This engagement may be lower than expected for the politics category - but this may be attributed to the overall plateau of politics with regards to COVID-19 over the past few months. In recent months, there have not been major changes in COVID-19 restrictions in the United States nor Canada, where most of the debate occurs.

We notice that seven of the top ten words are 'blasio', 'mayor', 'sector', 'mandate', 'employers', 'city, and 'nyc'. There was an abundance of tweets about the mayor of New York City, Bill de Blasio, and his new vaccine mandate for public sector workers [3]. 'Queensland' and 'reopen' were also among the top ten words; this aligns with the news that the Queensland border in Australia is reopening after a 141 day quarantine separating it from the rest of Australia [4]. Finally, 'trump' was also among the top ten words, referring to the former President of the United States. Interestingly, the current president, Joe Biden, did not appear, indicating much of President Trump's rhetoric and persona is continuing to circulate in conversations around COVID-19. Future work could contrast sentiment in tweets centered around Trump versus tweets centered around Biden to analyse vaccine sentiment differences in political bases in the United States.

Personal Life

The personal life category had a majority of negative sentiment (56%) tweets, with neutral (16%) and positive (28%) comprising the remaining tweets. Interestingly, this category contained the highest percentage of positive sentiment tweets out of all the topics. When compared with the measly 8% of positive tweets in the personal opinions category, this indicates that the broad opinions on the pandemic and vaccine a person might share are not necessarily grounded in the factual realities of their own lives. The personal events category reflects a reasonable balance between positive vs. negative sentiment (naturally skewed towards negative, as the tweets center around a global pandemic), while the personal opinions category is extremely weighted towards negative sentiments. This may reflect a bias towards sharing strongly negative opinions about the vaccine on Twitter.

This category was also among the top three in terms of engagement, accounting for 21% of the total tweets. The engagement in this category is logical, as many use Twitter as a platform to share anecdotes of their personal lives, especially in the context of COVID-19. Some may find it important to inform others of their vaccination status, if they have fallen ill from COVID-19, or if someone close to them is in the hospital due to severe COVID-19 illness. Among the top ten keywords were 'fever', and 'booster'. These tweets were centered around excitement of receiving or booking a booster vaccination appointment or discussed the side effects of the booster. Another keyword in the personal life category was 'smell', appearing in tweets discussing individuals' symptoms of COVID-19, and 'bed' reflecting the inability to get out of bed when ill. The keywords 'mom' and 'hotel' were also used to describe scenarios of loved ones falling ill, or of mandatory quarantine. Many of these were of negative sentiment, which can be expected, as we often complain and are upset when ill or isolated. The word 'lol' is a slang word used to say "laughing out loud", and appears frequently in positive sentiment tweets to describe happy feelings or share humorous stories or opinions. Keywords 'kody' and 'trash' refer to the television show *Sister Wives*, in which recent episodes contained drama related to COVID-19.

Workplace

The workplace category did not contain many tweets (22 tweets) but showed a relative balance between negative and neutral sentiment (32% and 50%, respectively), leaving 18% of the tweets with positive sentiment. Keywords in the top ten included 'investors', 'customers', 'company', 'businesses', and 'workers', indicating that this category narrowly captured business-related content. Keywords 'omg' and 'blessed' reflect the tweets about getting a job during COVID, after struggling to keep on at the beginning of the pandemic. Given the largely neutral sentiment of this category and the overall scarcity of work-related tweets in the sample, we believe the workplace is not a particularly relevant setting for discussions on COVID-19, the pandemic, and vaccinations. However, further samples would need to be collected to confirm this, particularly if and when companies choose to phase out work-from-home options, which would bring coworkers in close proximity more often. We must also consider the possibility of workplace tweets being underrepresented overall; employees may not wish to publicly express their opinions regarding their workplace, employers or co-workers in fear of losing their jobs. This reason may also account for the greater proportion of neutral sentiment tweets: employees that do choose to tweet about their work may decide to express themselves in a way that is more factual or matter-of-fact to ensure that they face no repercussions at work.

School

The school category had low engagement overall, with only 39 total tweets (3% of all tweets). The school category could also be under-represented as Twitter users between the ages of 13 and 17 and between 18 and 24 only account for 6.6%, and 17.1%, respectively, of all users in 2021 [5]. As these age groups make up the majority of school-aged children and adults, it is natural to have lower engagement in this category. The school category contained many school-related keywords like 'school', 'students', 'exams', 'class', and 'institutions'. Interestingly 'return' and 'diseases' also appeared in the ten most common words, which may reflect a broader conversation about the safety of children returning to classes after a year or more of virtual learning at home. The majority of tweets held negative sentiment (49%) while a sizeable minority were neutral (41%). Only 10% of tweets were positive. This could reflect parental angst about their children returning to class amidst the ongoing pandemic, especially since children only recently became eligible for

the vaccine [1]. The negative sentiment is also reflective of the frustration many college and university students have faced over the course of the semester. Some negative sentiment tweets came from parents complaining about their children's mask requirements at school, or expressed frustration with any new COVID-19 protocols that were implemented in schools. Some students disagreed with the protocols put into place by their institutions, or found these protocols to be confusing or contradictory. Both students and parents expressed concerns over the 'quality', of the education received. The few positive sentiment tweets represented the excitement and relief parents felt once the green-light was given to inoculate children. The majority of neutral tweets constituted announcements made by or about institutions to convey COVID-19 related information (e.g. protocols, vaccine updates).

Group Member Contributions

Emily wrote the code to explore the data, clean and filter the data, tokenize the data, compute tf-idf values for words in each topic, and collect sentiment data, and manually checked the word list used for tf-idf calculations. She plotted the number of tweets in each category, the number of tweets by sentiment, and the number of negative, neutral, and positive tweets in each category, and created the table with topic definitions and the top 10 words by tf-idf score for each topic. She also annotated 100 of the tweets, and wrote the Results section, and half of the Data section, and edited the rest of the sections.

Veronica wrote the code to access tweet content from the Twitter API and created the filter for keywords in the Twitter API. She helped Sonia create the categories and annotated a little under half of the dataset. She also wrote the Introduction section, half of the Data section, and half of the Discussion. She contributed to other sections by proof-reading.

Sonia wrote code to adapt Veronica's pulled tweets from a JSON format to a CSV format that was easier to annotate. She also developed the categories for annotation and annotated a little more than half the dataset. She wrote the Methods and half the Discussion and proof-read the report.

The three members communicated through a groupchat to discuss design decisions.

References

[1] FDA. 2021. FDA Authorizes Pfizer-BioNTech COVID-19 Vaccine for Emergency Use in Children 5 through 11 Years of Age. https://www.fda.gov/news-events/press-announcements/fda-authorizes-pfizer-biontech-covid-19-vaccine-emergency-use-children-5-through-11-years-age. Accessed: 2021-13-12.

[2] Kreier, F. 2021. 'Unprecedented achievement': who received the first billion COVID vaccinations?

[3] NYC Media. 2021. NYC to mandate Covid-19 vaccines for all private sector workers. https://www.cnn.com/videos/business/2021/12/06/nyc-mayor-bill-de-blasio-covid-19-vaccine-private-sector-mandate-vpx.nyc-media. Accessed: 2021-13-12.

[4] Queensland Government. 2021. Changes to Queensland's border restrictions at 13 December 2021. https://www.qld.gov.au/health/conditions/health-alerts/coronavirus-covid-19/current-status/queensland-restrictions-update/changes-to-queenslands-border-restrictions. Accessed: 2021-13-12.

[5] Statista. 2021. Distribution of Twitter users worldwide as of April 2021, by age group. https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users/. Accessed: 2021-13-12.

[6] The New York Times. 2021. UK Announces Omicron Cases and Extenends Booster to all. https://www.nytimes.com/2021/11/29/world/europe/uk-omicron.html. Accessed: 2021-13-12.

[7] WHO. 2021. WHO Coronavirus (COVID-19) Dashboard. https://covid19.who.int/. Accessed: 2021-13-12.