COVID-19 on Twitter: What Tweets Reveal About COVID in the Anglophone World

Brian Kehoe, Holly Sandvold, Huzbah Jagirdar
McGill University
brian.kehoe@mail.mcgill.ca, holly.sandvold@mail.mcgill.ca, huzbah.jagirdar@mail.mcgill.ca

Introduction

Twitter plays a key role in daily political and cultural discourse and over time it has come to be a vital tool to gauge public opinion. With the COVID-19 pandemic and the recent rollout of vaccinations for the virus all around the world, we used data from twitter to determine the conversations around these topics in the anglophone world. Our key objectives were to discover the main topics of discussion around COVID-19 and what they concern, determine the relative engagement around each issue; and find out the sentiments surrounding them.

To inform our analysis, we used data collected from the Twitter API to develop eight topics around the tweets and characterized them by computing the 10 words in each category with the highest term frequency - inverse document frequency (tf-idf) scores. We also coded whether the tweets had a negative, neutral or positive sentiment associated with them.

Our results were influenced by some key design decisions such as limiting the language to only english, using only hashtags to retrieve tweets, the timeframe within which we collected the data, and the topic coding method. Furthermore due to time constraints, we used single annotation instead of double annotation. We also ignored replies, hyperlinks and images attached to tweets in our annotation and analysis process.

We found that a majority of the current conversations around COVID-19 are in relation to vaccines and there is a fairly even split between the sentiments of tweets in this category. Additionally, overall, negative tweets around COVID-19 were observed equally as often as the positive ones. However, the sentiments observed in the other topics varied largely. Moreover, the top words computed using tf-idf scores gave us a satisfactory understanding of the discourse surrounding each of our developed categories. As a result we were able to make some insightful interpretations about COVID-19 and vaccines, around these conversations.

Data

Data Collection

Our data was collected using the python library tweepy to access the Twitter API. Originally, we attempted to use the get() method from the python requests library but we ran into issues with the API permissions. Ultimately, we found that the tweepy library was easier to use, more flexible, and improved code readability. In order to use the tweepy API, we first created a twitter developer account with basic permissions. We created a new project under our account in order to get a Bearer Token which would allow us to access the API. We used the search_recent_tweets tweepy function when collecting the tweets which returns tweets from the last seven days. We were interested in a three day time span - from the beginning of November 18, 2021 to the end of November 20, 2021 - so this function was sufficient. We narrowed our search by eliminating retweets and quote tweets from our request. We decided to keep only standard tweets and replies, although we later found that replies were difficult to interpret and were eliminated. We chose to search using the hashtag function. We were concerned that our topics, especially "covid" generally, would be referenced in passing in many tweets and we felt that the hashtags would give us the best chance of getting tweets intentionally targeting our topics of interest. We reasoned that users who put the time to include a hashtag had a greater investment in the topic than users who simply included a given word in their tweet. Since we were interested in collecting tweets related to coronavirus and the coronavirus vaccine, we searched for the following hashtags: #covid19, #covid, #coronavirus, #covidvaccine, #pfizer, #moderna, #johnsonandjohnson, #vaccine, #vax, and #vaccinated. We noticed that tweets referencing the name brands of vaccines used several variations of hashtags (for example, #johnson&johnson rather than #johnsonandjohnson) so in order to remain consistent, we decided to choose the most popular hashtag for each vaccine. This way, we targeted tweets about each vaccine equally.

Once we had prepared our search constraint, we collected tweets in batches of 100. The tweepy function limits the number of tweets collected per call to 100 and only allows a few hashtag specifications per call. In order to accommodate these parameters, we ran 20 calls to the API. First, we divided our timeframe of interest into 2 even blocks: 2021-11-18T00:00:00Z to 2021-11-19T12:00:00Z, and 2021-11-19T12:00:00Z to 2021-11-21T00:00:00Z. Then, in each block of time, we searched for each hashtag, one at a time. Using this method, we allowed for a maximum of 2000 tweets total to be collected, guaranteeing that we had the presence of every hashtag and that the tweets spanned the range of time we were interested in. Finally, we combined all the tweets. e kept only the tweets with a unique tweet ID, thereby eliminating any tweets we had gathered more than once by searching for one hashtag at a time. At this point, we had 1707 tweets in total.

Data Cleaning

Once we obtained our tweets, we began the process of cleaning the JSON file containing them. First, all the tweet entries that contained tweet replies were removed to make the process of annotating our data easier. Classifying tweet replies into a category and ascribing a sentiment to them would prove to be difficult since the annotators would be missing crucial context with regards to the tweet the response was to . Additionally, we removed all emojis and hyperlinks from the tweets. The Unicode emoji strings were removed using the string.encode() method with the parameters ‘ascii’ and ‘ignore’ and the string.decode() method. So, we encoded the whole tweet to ascii while ignoring the characters that could not be converted (the Unicode characters) and decoded it to its original format again. Hyperlinks were removed by identifying the patterns followed by links in tweets using the re library. We replaced them with the term “LINK” to help the annotators recognize that the tweet contained a hyperlink to another tweet or website. Moreover, all the newline escape sequences (“/n”) were also removed. We made the above changes during the data cleaning to increase the readability of the tweets and facilitate the annotation process. Lastly, 1000 tweets were randomly chosen and saved into a tab-separated values (TSV) file.

The final tweet data file contained the following fields: tweet_id, author_id, username, text, and created_at. The tweet_id provides a unique identification for each tweet. Similarly, author_id, and username identify unique users. Finally, text contains the tweets and created_at contains the timestamp.

Method

To begin, we collected one thousand tweets based on various language, content, and temporal filters discussed above. Once the tweets had been collected, we began the annotation process so the tweets could be analyzed for sentiment and topic. We first conducted an open coding of 200 tweets sampled at random. We coded each tweet with two values: category/topic and sentiment. To avoid any confusion caused by tweets that fell into several categories, we followed a hierarchy of coding. This hierarchy was decided on the basis of what topics would be considered the most important for the task at hand, which was to gauge the conversations around COVID and vaccines on social media. The top categories in the hierarchy cover vaccination and specific policies relating to pandemic response, while the lower priority categories cover broader areas in an attempt to capture general sentiment about how the pandemic has affected peoples’ health and lives.

We defined sentiment in a category-specific manner. To understand whether people are in support of or against vaccination, lockdowns, masks, and policies which are designed to slow the spread of coronavirus, we defined positive in policy-related categories to be “in favor of”, negative to be “against” and neutral to be “indifferent”. The remaining categories were coded using the following test: Does the author of the tweet like X? Where X is the topic of the tweet. This way we capture general sentiment about the topics, and not feelings towards any one policy.

Before dividing up the remaining tweets amongst the group members, we decided to annotate 50 of the same tweets and calculate a Fleiss Kappa statistic, which measures agreement between more than two annotators (Fleiss 1971), as a proxy for how consistent our annotations would be on the larger tweet sample. The formula for a fleiss kappa statistic is the agreement achieved between annotators minus the agreement expected from chance over one minus the agreement expected from chance. In other words, the statistic relates how much actual, non-chance agreement occurred divided by the maximum possible agreement that cannot be explained by chance. We achieved a fleiss kappa score between 0.81 and 1 for category annotation, which indicates near perfect agreement amongst annotators, and between 0.61 and 0.8 for combined category-sentiment annotation, which indicates substantial agreement (Landis and Koch 1977). These scores were satisfactory, but as an additional measure of quality assurance we agreed to independently annotate the remaining tweets in chunks of 10 minutes to avoid generating errors from fatigue.

Once all 1000 tweets had been annotated, we calculated tf-idf (term frequency-inverse document frequency) scores for the words in each category. The tf-idf score indicates how important a given word is to characterize a category - the greater the score, the greater the importance of that word. To ensure relevance of the words we analyzed, we removed words which appeared less than five times in the document and default English ‘stop words’[1] such as “and”, “the”, “at”, and so on. Additionally, as part of the preparation before calculating tf-idf, all numbers and punctuation except ‘#’ were removed and the text was made lowercase. This prevents covid-19, covid, and COVID from counting separately, for example. The calculation of tf-idf scores was conducted in two steps: First, the total frequency (tf) of words by each category was calculated and saved in a JSON file. Next, inverse document frequency (idf) was calculated by taking the log of the quotient of the total number of categories over the number of categories in which a given word appears. The product of these two values represents the tf-idf score, which was then used to rank the words by category in descending order of importance so the top 10 could be collected.

Results

Using our final eight topics - Vaccine, Lockdown, Masks, Economic, Political, Health, Cultural, and Other - and our three sentiment options - positive, neutral, and negative - we assembled the data on tweets per topics, total sentiments per topics, and the words with the highest tf-idf scores per topic. We are confident in the consistency of our annotations due to the fleiss kappascores of 0.81 and 0.69 which we received for category-only assignment (i.e.: Vaccine) and category-sentiment combined assignment (i.e.: Vaccine,positive), respectively, for the 50 tweet samples which we collectively annotated.

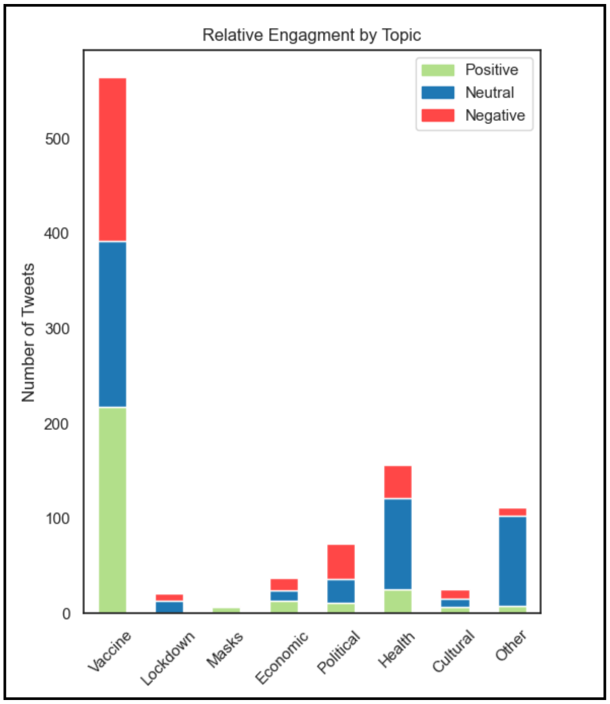


Figure 1: Relative Engagement by Topic

We found the following total number of tweets per topic: 565 Vaccine, 21 Lockdown, 8 Masks, 37 Economic, 74 Political, 156 Health, 26 Cultural, and 113 Other. (Figure 1) As we were targeting vaccine-related tweets with both our hashtags and our hierarchy of identification, it is consistent that the Vaccine topic had by far the largest number of tweets.

The overall sentiments of the tweets were 43% neutral, 29% positive, and 29% negative (Figure 2). The sentiments per category varied largely (Figure 3). Sentiments in the Vaccine, Cultural and Economic categories were fairly evenly split while all of the other categories had one dominating sentiment. Lockdown tweets had no positive sentiments and 62% neutral sentiments. Mask tweets had no negative sentiments and 88% positive sentiments. The Health and Other categories both had a majority of neutral tweets while the Political category had a majority of negative tweets.

We can improve our understanding of the content of the tweets by looking at the top words by tf-idf scores by topic (Figure 4). In the Vaccine category, the top ten words largely referenced booster shots as well as “slots” and “doses”. In the Lockdown topic, the country Austria came up often, as well as the terms “march” and “freedom”. The word “spread” had the largest score in the Mask topic, followed by “cdc”, “days”, and “pandemic”. The Economic topic featured “million”

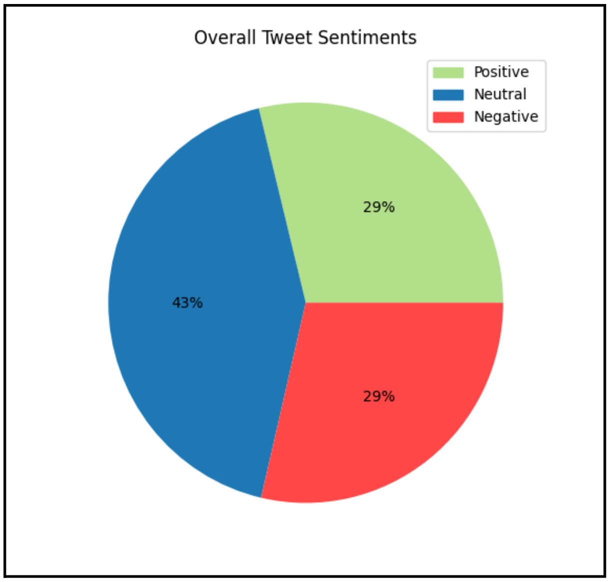


Figure 2: Overall Tweet Sentiments

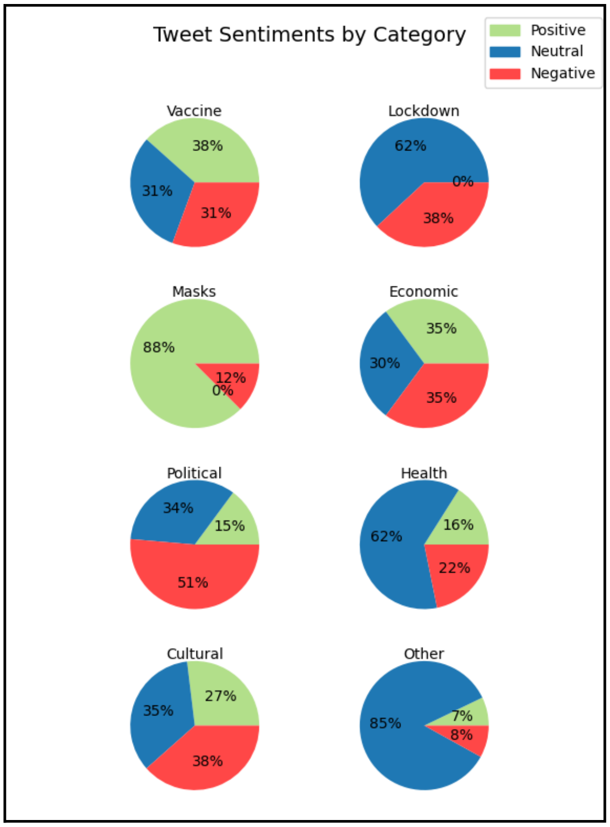


Figure 3: Tweet Sentiments by Category

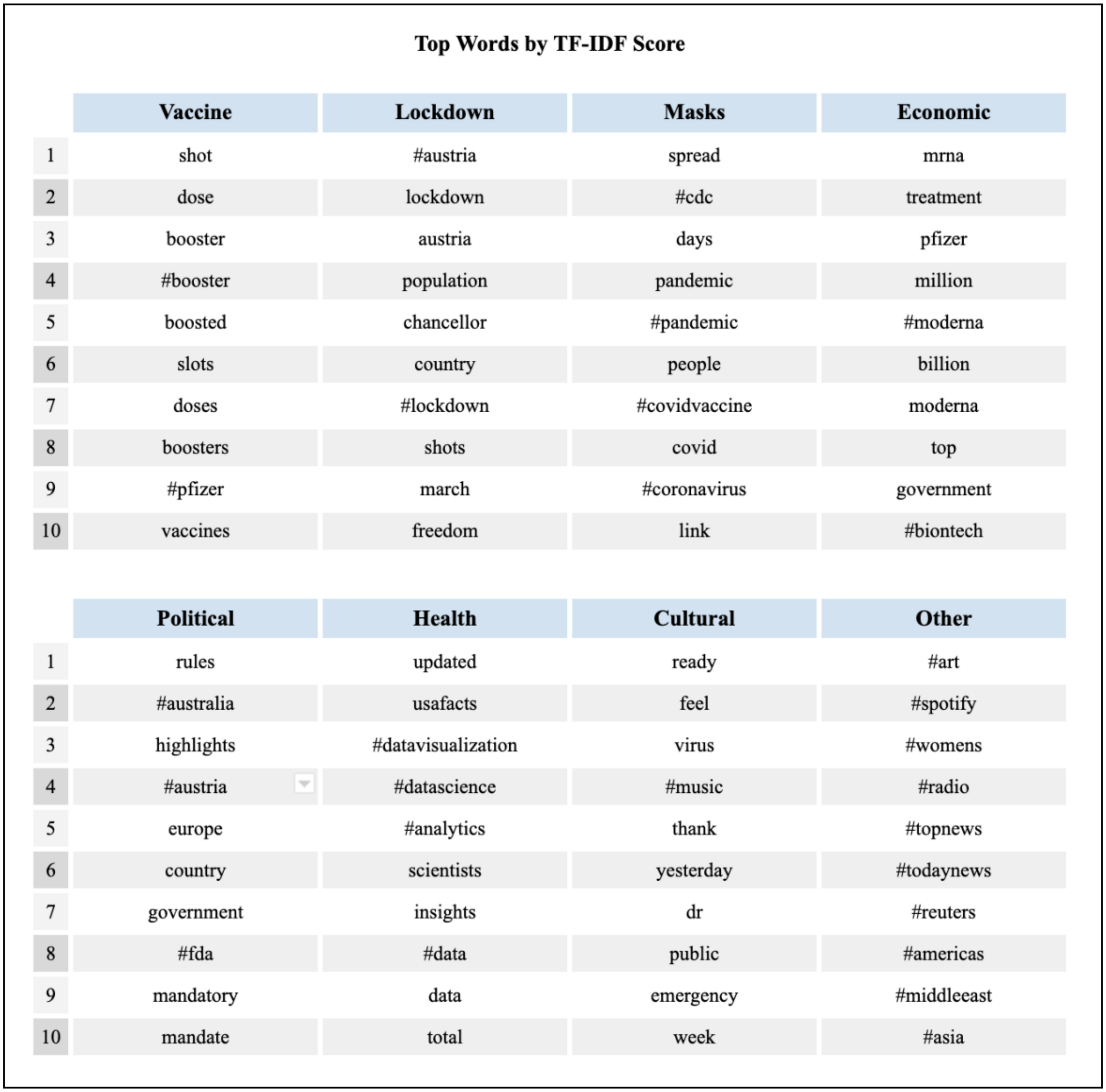
and "billion", as well as several mentions of vaccine name brands like "pfizer" and "moderna". The Political category again mentioned Austria, as well as Australia. It also featured the terms "rules", "mandate", and "mandatory". The Health category was dominated by quantitative terms like "data", "usafacts", and "#analytics". The Cultural topic had a range of words from "dr", "virus" and "emergency" to "#music" and "feel". Lastly, the Other category's top words were all hashtags like "#art", "#spotify", and "#radio".

While we had several topics with a large number of tweets to analyse, like Vaccine, Health, and Other, some of the topics only contained a handful of tweets. The Mask, Lockdown, Cultural, and Economic topics had small sample sizes and therefore further research is required to determine the public sentiment on these topics. In addition, the top word scores in these four categories carry less weight due to their small sample sizes.

Discussion

Keeping in mind the limitations of our methodology, we interpret the sentiments and the top words in each category and make recommendations based on our analyses.

The words in each category indicate that the tweets capture public discussion around a topic at a particular moment in time. The tweets were sampled at a time when Austria announced a national lockdown, booster shots were released, and there was concern of a COVID wave spreading through Europe. These events are mirrored in the top words of the Lockdown, Vaccine, and Political categories respectively. This provides confidence that the collected tweets captured the most popular discussions of the moment. That being said, because the tweets were collected in a narrow window of time, sentiment regarding particular events could have overpowered broader sentiment on a topic. A further limitation is that only English language tweets were selected. This ignores tweets made in other languages within Canada, or any other country of interest, and also does not guarantee that the tweet will pertain to a predominantly english-speaking country. Some tweets collected were about state government meetings in Rajasthan, for example, which may not be relevant to a Canadian non-profit trying to understand covid sentiment amongst English speakers. We chose to collect tweets by hashtag alone, which resulted in an Other category of 113 irrelevant tweets. Using keywords within tweet text along with hashtags may reduce the size of this category in later research. The final salient limitation in our research was single annotation. Future research should include a double annotation in order to more accurately categorize tweets



Top Words by TF-IDF Score

	Vaccine	Lockdown	Masks	Economic
1	shot	#austria	spread	mrna
2	dose	lockdown	#cdc	treatment
3	booster	austria	days	pfizer
4	#booster	population	pandemic	million
5	boosted	chancellor	#pandemic	#moderna
6	slots	country	people	billion
7	doses	#lockdown	#covidvaccine	moderna
8	boosters	shots	covid	top
9	#pfizer	march	#coronavirus	government
10	vaccines	freedom	link	#biontech

	Political	Health	Cultural	Other
1	rules	updated	ready	#art
2	#australia	usafacts	feel	#spotify
3	highlights	#datavisualization	virus	#womens
4	#austria	#datascience	#music	#radio
5	europe	#analytics	thank	#topnews
6	country	scientists	yesterday	#todaynews
7	government	insights	dr	#reuters
8	#fda	#data	public	#americas
9	mandatory	data	emergency	#middleeast
10	mandate	total	week	#asia

Figure 4: Top Words by TF-IDF Score

which could often be grouped under several of the defined topics.

With those caveats in mind, certain trends are still apparent. Vaccine sentiment was 69% neutral or positive. Neutral for the Vaccine topic includes informational tweets which, while not ostensibly positive, are not roadblocks to having a more vaccinated, or at least adequately informed, general public. The Health category is similar to Vaccine in this way. While 16% positive appears low, 78% of tweets are positive *or* neutral. The interpretation of this is that a clear majority of tweets in this category either shared neutral public health information which was designed to slow the spread of the virus, discussed new non-vaccine covid drugs, or told positive stories of recovery or improving public health conditions. The Mask category showed an overwhelmingly positive sentiment, but only contained 8 tweets. As a result, the tf-idf scores and sentiment must be interpreted cautiously. The first few words could could reasonably be interpreted as relatively unique to the topic but “covid” and “#covidvaccine” assuredly appeared often in the other topics.

The Lockdown, Political, Economic, and Cultural categories suggested a less positive discussion regarding the pandemic. Lockdowns received zero positive tweets; all Lockdown tweets were informative or negative. This may be skewed by Austria’s recent unpopular choice to lockdown after a long period of being reopened, rather than sentiment related to lockdowns in general. Political tweets predominantly referred negatively to mandatory covid policies, allegations of government corruption, and Australia’s recent decision to enforce isolation measures. There is a possibility that this category is overrepresented by complaints rather than praise due to our coding system. Tweets which praised policies which are helping to end the pandemic would have been categorized as health-related if health was mentioned in the tweet, like this tweet thanking the government of Halifax for its well-designed testing centers: “another fantastic experience getting a rapid #Covid test at #Halifax convention centre”. Economic tweets were split evenly between the three kinds of sentiment, but neutral in the economic category tended to be financial reporting, not information which could have positive ancillary effects on the public like the neutral vaccine tweets. Positive economic tweets tended to describe post-covid recovery or the personal gain of individuals, while negative tweets tended to describe pandemic-profiteering by large pharmaceutical companies, which may erode trust in the vaccine or nascent covid medication in the population. Negative economic tweets are very important to track as negative economic conditions can lay the groundwork for social unrest and lack of faith in public institutions. The Cultural category had a substantial number of negative or neutral tweets. This suggests that individuals were generally dissatisfied with the cultural consequences of covid, like increased virtual interaction or anti-asian racism. Neutral in this category includes factual updates about an event or a relevant cultural figure - like stating that a celebrity is isolating after contracting the coronavirus. This category had several words in its top scores that we expect are not necessarily unique to the topic, like “virus” and ‘dr”. This does provide insight on the dual nature of covid as an epidemic and as a cultural event.

Tweets in the Other category were, by definition, not relevant to the pandemic in any capacity. The top 10 most frequent words in this category were hashtags because these tweets tended to be advertisements which included some of the most popular hashtags on twitter in an attempt to reach more users.

The results gathered demonstrate that when it comes to discussions which relate directly to health, like vaccines, new drugs, or cases, tweets tend to be more positive. Vaccine hesitancy, specifically, did not commonly appear within the vaccine category. A possible interpretation of this is that the negative vaccine tweets were made by people decidedly against vaccination versus being unsure about its effects. Encouragingly, the recently deployed booster shot was being discussed on twitter with considerably more positive sentiment than negative. This bodes well for the reception of future vaccination measures which may be implemented. The vaccine topic had significant engagement as defined by the number of tweets, but further research could investigate engagement as the number of likes or retweets of tweets within a given category to have a clearer understanding of how likely a certain tweet is to spark additional conversation. Discussions surrounding policies that are intended to stop the spread by imposing more serious restrictions on people’s behavior, like lockdowns or masks, were met with more negativity. It is indeed telling that one of the top words in Lockdown is “freedom”.

Overall, vaccine sentiment within the tweets collected suggests that the conversation is mostly informational or positive, which provides an encouraging outlook for institutions which are concerned about waning trust in the population regarding vaccines. The strong negative reaction to lockdowns and related policies clearly shows that the public is resistant to further isolation, which is a key insight for policymakers when deciding how to balance public safety measures with morale and compliance.

Group Member contributions

Brian annotated one third of the final tweets, including 200 for the initial open coding, annotated 50 tweets used for the Fleiss Kappa statistic, designed the coding scheme, calculated the Fleiss Kappa statistic and the tf-idf scores, and wrote those sections of the paper accordingly. He, along with the other members of the group, planned meetings and delegated responsibility.

Holly collected the tweets using the Twitter API, created the graphs and tables using Matplotlib and

Seaborn, and wrote about Data Collection and Results. In addition, she annotated a third of the tweets, helped create a timeline of deadlines and tasks for the project, and participated in meetings with the group.

Huzbah cleaned the collected data, and helped with the calculation of the tf-idf score by computing the term frequency and making other key design decisions. She annotated a third of the tweets and participated in meetings. Additionally, she wrote sections of the paper relating to the method, data and introduction.

Notes

[1] gist.githubusercontent.com/larsyencken/1440509/raw/53273c6c202b35ef00194d06751d8ef630e53df2/stopwords.txtl

References

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5): 378-382. https://doi.org/10.1037/h0031619.

Landis, J. R.; and Koch, G. G. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33(1): 159-174. https://doi.org/10.2307/2529310

Twitter. 2021. Search Tweets: GET /2/tweets/search/recent. https://developer.twitter.com/en/docs/twitter-api/tweets/search/api-reference/get-tweets-search-recent. Accessed: 2021-11-22.