A Twitter-Based Analysis of the Perception of Covid-19 in North America

Neeraj Katiyar, Cesare Spinoso-Di Piano, Ziming Wang
McGill University, Montreal, Canada
845 Rue Sherbrooke O
Montreal, Quebec, Canada
{neeraj.katiyar,cesare.spinoso-dipiano,ziming.wang2}@mail.mcgill.ca

Introduction

Covid-19 has become one of the greatest threats to human beings in terms of health, society, and economy. Up to this moment, there have been little to no signs of remission. During this period, social media platforms like Twitter have played an important role in spreading information about the pandemic across the world. Moreover, the use of social media as a way for people to express their opinions and feelings on different topics has progressively expanded. As a result, analyzing covid-related tweets can provide significant insights in understanding the views and opinions of people in our society regarding the pandemic.

Our work includes collecting covid-related tweets across North America, manually annotating the category and sentiment for each tweet, and computing term frequency–inverse document frequency (tf-idf) scores for top words in each category. This work concentrates on extracting categories and sentiments related to COVID-19 on Twitter and analyzing this data to provide insights about the public perception of the pandemic and other pandemic-related topics such as vaccination. The key findings of our research are summarized as follows:

1. There is a significant proportion of the population that remains apprehensive regarding the effectiveness of vaccines and public health measures such as vaccine and mask mandates.
2. There are signs that some aspects of society have started to recover from the pandemic. Our analysis shows that social life and economic trends have begun to return to a pre-pandemic state.
3. In contrast with the previous point, there are also signs of growing concern regarding the covid variant first discovered in South Africa. This has contributed to a general feeling of anxiety and of dread in the population regarding the possibility of the pandemic persisting longer than expected.

Data

We collected tweets from the Twitter API over the course of three days from November 23rd to November 25th. We pulled 1000 tweets for each day at roughly the same time (between 7PM and 10 PM Eastern Standard Time). To collect these tweets, we used keywords related to the coronavirus including “covid”, “corona”, “coronavirus” and “virus” as well as keywords related to the vaccine such as “vaccine”, “vaccination” and the names of the most prominent covid vaccines including “pfizer”, “biontech”, “johnson & johnson” and “astrazeneca”. Moreover, to keep tweets only from North America we used longitude-latitude-radius filters to approximate the region. To do so, we used an online visualization tool (Free Map Tools 2021) which allowed us to visualize the coverage of each geolocation filter. As a result, the scraped tweets consisted mostly of Canadian and American content. In addition, we leveraged Twitter’s API to only keep tweets that were written in English and were not retweets. Some additional filtering was applied to each set of 1000 tweets to eliminate duplicate tweets (i.e., tweets with the same text) as well as replies (using the `in_reply_to` fields). This led to dataset sizes of 540, 509 and 583 for the 23rd, 24th and 25th of November respectively.

Method

As described in the data section, we collected 1000 tweets for each of the 3 days between November 23rd and November 25th. We collected such a large volume because we were expecting filtering operations such as the removal of replies and retweets to shrink our initial dataset size. In addition, an importance was placed on fetching the tweets at approximately the same time. This is because in addition to uncovering the most salient topics discussed surrounding covid we were also interested in explaining their possible change through time.

For our annotation procedure, we annotated the first 65 tweets for each day as a group to make sure that the topics we were creating were well defined. Our approach at annotation started with using very specific topics for each tweet on a first pass and clustering the similar topics into larger categories until we had 8 topics. To ensure that each topic was as informative as possible, we decided not to include an “Others” category as we believed this would not provide us with useful insights. Instead, based on the 8 topics that we had defined to tag the first 200 tweets, we decided only to annotate tweets that fell within those topics and skip any other tweets. While acknowledging that this may have led to a loss of information, we believed that the categories we selected were informative enough to provide insights about discussions related to the pandemic and vaccines. For sentiment annotation, we annotated tweets as positive or negative only when there was a clear first-person sentiment being expressed by the tweet. Otherwise, the tweet was annotated with a neutral sentiment even if the information it was conveying might have been perceived as having some sentiment. The process of annotation was split equally among the three team members and a final check for each annotation was done by one team member to ensure that annotation remained consistent throughout the 3 days of extracted data. The annotation procedure produced a total of 1024 tagged tweets.

To analyze our annotated data, we grouped tweets by their corresponding topic and computed the 10 words with the highest tf-idf scores. On a first pass, we noticed that all top 10 words were either stop words or punctuations which motivated their removal from the computation of the scores. Moreover, we lowercased the words as we noticed that terms such as "Biden" and "biden" were occurring in the top 10 without providing additional information. Thus, our tf-idf scores were computed based on the lowercased alphabetic words of the collected tweets filtering out stopwords using the nltk package (Bird, Klein, and Loper 2009). To analyze the sentiment of these discussions, we also computed the distribution of the neutral, positive, and negative class for each topic. Finally, as we noticed a significant topic change in our data throughout the 3 days, we also grouped the frequency of each topic by the days of the data extraction to explore this further.

TF-IDF Analysis

To signify the importance of each word in the related categories, we computed a tf-idf score for each word in tweets. We used the following version of the formula to compute the scores.

$$\text{TF-IDF}(w, c) = \text{TF}(w, c) \times \text{IDF}(w)$$

Where the left-hand side of the product is defined as

$$\text{TF}(w, c) = \text{the count of the word } w \text{ in the category } c$$

And the right-hand side of the product is defined as:

$$\text{IDF}(w) = \log\left(\frac{\text{the total number of categories}}{\text{the number of categories containing word } w}\right)$$

Results

After our initial open coding, we devised the following topics related to the pandemic and the vaccine:

- **Vaccine effectiveness:** Any mention of the covid vaccine's effectiveness at preventing infection. We also included tweets refuting the vaccine's effectiveness and suggesting its side-effects. This category was also defined to be the primary proxy for measuring vaccine hesitancy in North America (the sentiment and keywords of the other categories can be considered as weaker proxies for measuring hesitancy).
- **Pandemic statistics:** Any tweet providing statistics or numbers pertaining to the coronavirus and the vaccine. For example, the percentage of vaccinated people or the number of cases in a day. This class was mostly objective and thus observed low non-neutral sentimentality.
- **Pandemic social life:** Any tweet mentioning a social activity in the context of the pandemic. For instance, not being able to spend Thanksgiving with one's family because of a Covid outbreak.
- **Pandemic politics:** Any tweet mentioning a political figure or stating a politically grounded statement related to the pandemic. For example, tweets commenting on the difference in which the pandemic was handled between the Trump and Biden administration.
- **Pandemic mitigation measures:** Any tweet mentioning measures taken at either a local or federal level to mitigate the impact of the pandemic on public health. For instance, the mandating of vaccines or the rollout of vaccines for children.
- **Economic consequences:** Any tweet mentioning the economic impacts of the pandemic both at a local level (e.g., the shutting down of small businesses due to a lack of government funding) and at a global level (e.g., stock market fluctuations in response to the pandemic).
- **Covid variant:** Any mention of a mutation of the coronavirus including the "South African" variant (at the time of the extraction this strain had yet to be named).
- **Covid symptoms:** Any tweet about a user's covid symptoms or the covid symptoms of others.

We produce the distribution of the most salient topics over each day (Table 1) as well as the most frequent words for each topic (Table 3). At a high-level, we can observe that some of the results are skewed towards the events that were occurring during the time period of extraction. That is, between the 23rd and 25th of November two prominent events, the approach of American Thanksgiving and the emergence of a new covid variant, occupied most of the talking points in the North American Twitter landscape. These form the most notable "outliers" of our results. The distribution of sentiment for each topic (Figure 1 and Table 2) is measured on an aggregate level for compactness. Other than with the covid variant topic, we did not observe a significant shift in the sentiment of topics across different days as can be seen in the daily sentiment bar plots (Figure 2, 3, 4 found in Appendix A).

In addition to the shift in distribution caused by Thanksgiving and the new covid variant, the terms "aaron" and "rodgers" appear to be the only other noticeable outliers in our tf-idf table. The reason that the terms "aaron" and "rodgers" are top words in the topic of "Covid Symptoms" is because Green Bay quarterback Aaron Rodgers had been dealing with a toe-related injury and many speculated that it was "COVID toe". "COVID toe" is a condition that can cause skin lesions on a person's toe after a COVID-19 infection which can last several months.

Date	Topics								
	Pandemic statistics	Pandemic mitigation measures	Pandemic social life	Vaccine effectiveness	Covid variant	Pandemic politics	Covid symptoms	Economic consequences	Total
2021-11-23	58	82	49	67	5	50	21	11	343
2021-11-24	42	63	51	75	4	38	38	13	324
2021-11-25	28	54	49	51	95	32	26	22	357
Total	128	199	149	193	104	120	85	46	1024

Table 1: Table of tweet counts by date and topic

Topics	Sentiment			
	Negative	Neutral	Positive	Total
Pandemic statistics	12	116	0	128
Pandemic mitigation measures	37	148	14	199
Pandemic social life	33	79	37	149
Vaccine effectiveness	39	108	46	193
Covid variant	19	85	0	104
Pandemic politics	44	75	1	120
Covid symptoms	27	54	4	85
Economic consequences	7	32	7	46
All	218	697	109	1024

Table 2: Table of tweet counts by topic and sentiment

Discussion

We explore the main takeaways and insights from our data in the following sections.

Vaccine hesitancy: There is a split in sentiment among people regarding the vaccine and its ability to end the pandemic. On the one hand, there are people who see the vaccine as safe and effective demonstrated by the positive sentiment of the “Vaccine effectiveness” topic and the use of terms such as “safe”. On the other hand, people who are vaccine hesitant or seem to see the vaccine in a negative light appear to be primarily concerned with its adverse effects as demonstrated by the presence of the term “adverse” in this category. This negative position towards the vaccine is paralleled by the overall negative perspective of mitigation measures such as vaccination passports and vaccine mandates. This is evidenced by the fact that the number of tweets with a negative sentiment towards mitigation measures (37) appear to be twice the number of those with a positive sentiment (14) and some of the most frequently occurring words include “passport” and “mandate”. Finally, this negative sentiment towards the vaccine and its ramifications to other parts of society (e.g., public health mandates) appear to have become a very politicized issue. The fact that almost all sentiment regarding “Pandemic politics” is negative as well as the presence of the term “mandate” as one of the most frequent words in this category appear to support this. Furthermore, though this is only speculation, the fact that the term “conservatives” appears in the most frequent words and that “liberals” does not may also suggest that most anti-vaccine and anti-measure tweets involve a more conservative-leaning population. This is consistent with a recent USA-based study (Marc Debus 2021) which suggests that people who “identify as conservative express less intent to be vaccinated than individuals who identify with a different ideology.”

Pandemic persistence: Another key takeaway that resonates from our results is that two years after the beginning of the pandemic, the coronavirus has become a topic of discussion in every aspect of our lives. Indeed, from the way we socialize to how we establish and prioritize policies, the coronavirus has become omnipresent in our lives. This is demonstrated not only by the breadth of the topics that we extracted, but also by the volume of tweets that we were able to gather (1000 in the span of less than 3 hours for each day) given our restrictive constraints. Though this observation may seem obvious, it shows that the initial estimates of experts for a return to normalcy were wrong. Indeed, the fact that the most frequently occurring topics were “Pandemic mitigation measures” and “Pandemic statistics” rather than a topic related to “Pandemic recovery” shows that the coronavirus is still actively spreading and affecting public health. The persistence of the pandemic and the constant rate of new cases also explains why “Covid symptoms” was another prominent category in our findings. The fact that “deaths” and “hospitalized” were the most frequent words in the “Pandemic statistics” and “Covid symptoms” category respectively also show that the severity of the virus has not diminished. Thus, estimates by leading scientists such as Dr. Anthony Fauci (Alvin Powell 2020) of “normality by the end of 2021” did not account for extenuating circumstances such as a high levels of vaccine hesitancy and, as a result, were incorrect.

Recovery from the pandemic: While vaccine hesitancy as well as the persistence of the pandemic are negative trends that were extracted from the tweets, there are also some more positive aspects related to the pandemic that sur-

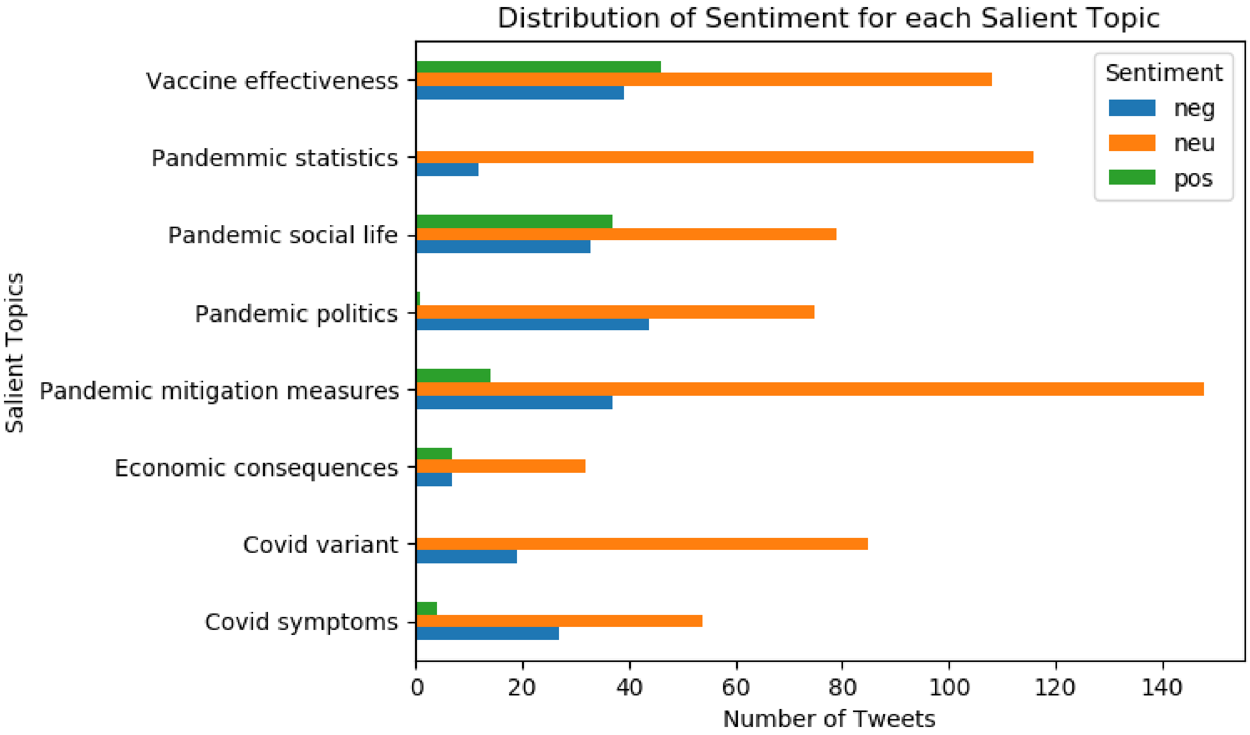

Figure 1: Bar plot of the distribution of tweet counts per topic and per sentiment

Topics	Top words
Pandemic statistics	deaths, reported, insights, analytics, usafacts, active, reports, confirmed, eight, recorded
Pandemic mitigation measures	require, mandate, law, employees, federal, passport, clinic, company, border, testing
Pandemic social life	family, dinner, game, test, mom, loved, seeing, tickets, saving, friends
Vaccine effectiveness	booster, shot, safe, moderna, pfizer, dose, adverse, cut, effects, flu
Covid variant	variant, mutations, scientists, africa, south, special, identified, concern, travel, strain
Pandemic politics	biden, trump, mandate, opposed, conservatives, court, unnecessary, vote, cuomo, navarro
Covid symptoms	taste, actually, foot, toe, adams, smell, symptoms, hospitalized, aaron, rodgers
Economic consequences	market, markets, futures, dow, stock, rent, amid, thrive, gold, asian

Table 3: Words with the highest tf-idf score per topic in descending order (from left to right)

faced from our analyses. The category of “Pandemic social life” was the only topic (other than “Vaccine effectiveness”) where there were more positive tweets (37) than negative tweets (33). As many tweets mentioned Thanksgiving (this term did not appear in the top 10 words because it was mentioned across most categories), this appears to be related to the fact that, unlike last year’s Thanksgiving, more people were getting together with their friends and their family. This is demonstrated by the key words in this category which include “family”, “friend”, “dinner” and “game”. We can also speculate that the users who provided positive tweets regarding social life may also have a positive view of the vaccine and its potential to allow a return to normalcy, but this conclusion cannot be made from our results as there is no striking overlap between these two topics. Further data scraping and analyses are needed to investigate this hypothesis. Though a majority of the tweets related to social life are positive and reflect the first steps of a return to normalcy, there is also a significant number of negative tweets. This may be explained by the fact that, although these gatherings are allowed, there still may be some concern over the safety of these activities. The prominence of “test” in this category aligns with reports that Americans across the country stocked up on PCR tests this Thanksgiving (Katherine J. Wu 2021). Thus, negative sentiment may be the result of not being able to gather for the holidays because of a positive test. In addition to the social aspect, the equal number of positive and negative tweets in the “Economic consequences” topic as well as the presence of the word “thrive” (and the absence of any negative terms) show that world economies and markets are very slowly recovering from the pandemic and even starting to thrive again.

Concern over the variant: Another unexpected takeaway from our data extraction and analysis was the spike in concern over the new coronavirus variant. As we can observe from the distribution of tweets per extracted day, any mention of a coronavirus mutation and of the covid variant first discovered in South Africa was practically non-existent on the first two days of extraction. The spike in the mention of the covid variant coincided with the first breaking reports of its discovery. In fact, travel restrictions were applied the 26th (Wang T. 2021), one day after our final extraction date. The spike accounted for close to a third of the tweets extracted on November 25th (95 out of 357) which is the largest proportion of mentions of a topic across all three days of data. This shows that the degree of concern was higher than in any of the other topics possibly due to its implications on eventual new lockdown orders. As expected, all sentiment regarding the variant is negative showing that the dread and fear of the pandemic persisting is universal regardless of political ideologies or views regarding the vaccine.

General anxiety and people's opinion towards the pandemic: Finally, on a higher-level note, we observe a general level of discontent and anxiety in the population regarding the pandemic. Only 10.6% of tweets (109) are labeled as positive whereas twice as many (21.2% or 218) are labeled as negative. In other words, most people expressing a sentiment regarding the pandemic (exactly 67%) hold negative attitudes towards the current situation. This shows that the population in general maintains a high level of anxiety and dissatisfaction towards the way in which the coronavirus has been handled. Furthermore, we also observe that 36.7% of tweets in the "Pandemic Politics" category are annotated as negative, which is the highest negative tweets percentage among all the collected categories. Given that the top 2 most frequent words in the "Pandemic Politics" category are "biden" and "trump", this may suggest that people are most unsatisfied with their government's policies surrounding COVID-19, especially in the United States. This reinforces the previous point about mitigation measures and suggests that governments and policymakers need to make more effective policies that reflect the needs of the entire population.

References

Alvin Powell. 2020. Fauci says herd immunity possible by fall, 'normality' by end of 2021. https://news.harvard.edu/gazette/story/2020/12/anthony-fauci-offers-a-timeline-for-ending-covid-19-pandemic/. Accessed: 2020-12-06.

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Free Map Tools. 2021. Free Map Tools. https://www.freemaptools.com/radius-around-point.html. Accessed: 2021-12-11.

Katherine J. Wu. 2021. The One Thanksgiving Necessity America Forgot to Stock. https://www.theatlantic.com/health/archive/2021/11/coronavirus-tests-thanksgiving/620663/. Accessed: 2021-11-10.

Marc Debus, J. T. 2021. Political ideology and vaccination willingness: implications for policy design. *Nature Public Health Emergency Collection*, 20(1): 1–15.

Wang T. 2021. United States will bar travellers from 8 countries in Southern Africa. https://www.nytimes.com/2021/11/26/world/americas/us-travel-restrictions-new-covid-variant-omicron.html. Accessed: 2021-12-05.

Appendix A

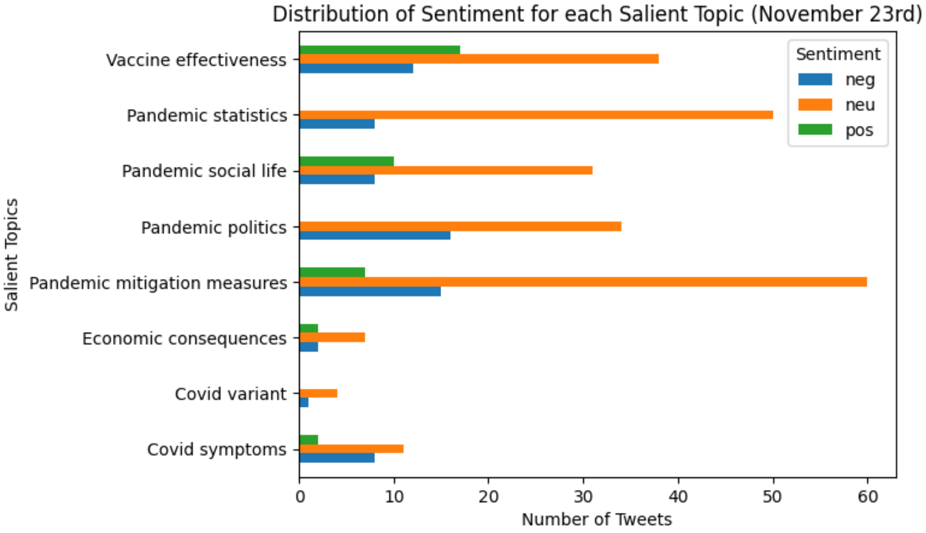

Figure 2: Bar plot of the distribution of tweet counts per topic and per sentiment for November 23rd

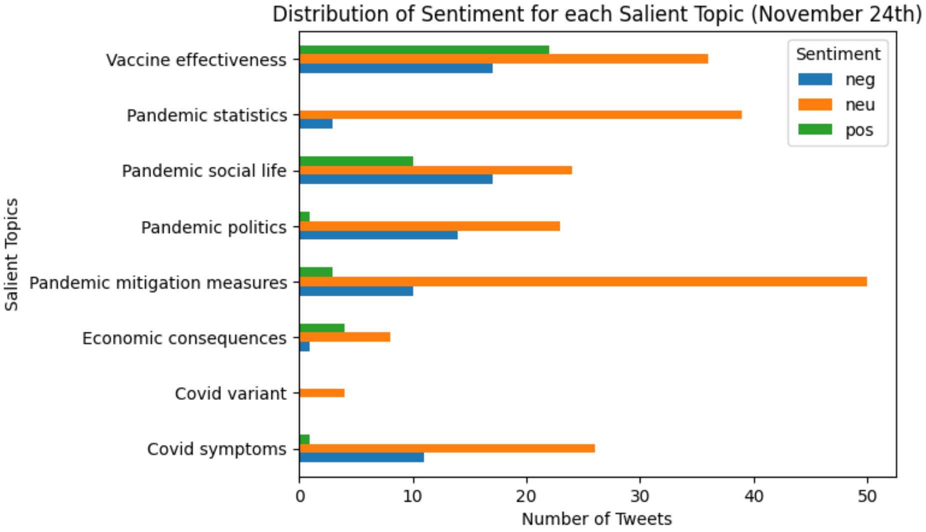

Figure 3: Bar plot of the distribution of tweet counts per topic and per sentiment for November 24th

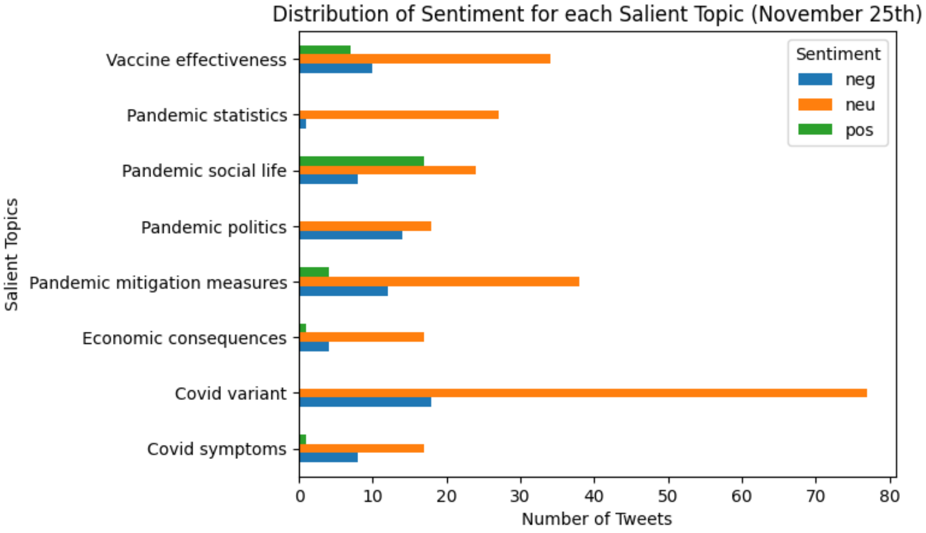

Figure 4: Bar plot of the distribution of tweet counts per topic and per sentiment for November 25th