DS 804: Exploration and Communication of Data
Final Project Report
Hannah Wirth, Dikshant Joshi, Lisa Olsson, Rajashekar Reddy Vemula

TWITTER SENTIMENT ANALYSIS – COVID-19 VACCINATIONS

Introduction and Overview

The objective of this project is to analyze the sentiment surrounding COVID-19 vaccinations now that vaccines have become widely available for consumers. To conduct this analysis, we followed a modified methodology from the "COVID-19 Vaccination Awareness and Aftermath: Public Sentiment Analysis on Twitter Data and Vaccinated Population Prediction in the USA" by Naw Safrin Sattar and Shaikh Arifuzzaman. Following our analysis, we compared our results to Sattar and Arifuzzaman's results to determine the differences in public sentiment from when COVID-19 vaccinations were first released to the present.

There are a few differences in methodology that should be noted:

1. Rather than Python, we used R and RStudio to complete our Twitter scraping.
2. We analyzed a total of 186,452 tweets while the original study analyzed collected 1.2 million tweets.
3. We collected tweets from an 11-day range while the original study collected tweets from a 5-week range.
4. We used the afinn sentiment analysis package rather than the TextBlob and VADER packages
5. We did not retrieve any location data for the tweets collected
6. We did not conduct any time series forecasting models.

For the second part of our project, we analyzed a variety of data sources relating to COVID-19 vaccinations and how they relate to the distribution of vaccinations, the number of cases, variants, and general sentiment about vaccine hesitancy.

Methodology

Computational Tools and Libraries Used

For the sentiment analysis portion of our project, we used R and RStudio to complete our sentiment analysis. To collect Twitter data and scrape tweets, we used the rtweet package in R. To obtain the sentiment scores of each tweet, we used the afinn package in R.

To create graphs to analyze our Twitter data and sentiment scores, we utilized Power BI to create a variety of visualizations.

To complete our analysis of COVID-19 vaccinations across the world and the United States, we used a variety of datasets compiled from Github, the CDC, Household Pulse Survey and Kaggle. All of these datasets were combined to create a unique data story to answer the following questions:

1. What has the diffusion of vaccines across the world looked like?

2. What is the progress in vaccination rates for each state over time?
3. What has happened to the number of cases as the vaccination rate has increased?
4. What is the relationship between vaccinations, new variants, and the number of cases?
5. How has COVID-19 vaccine hesitancy evolved across the US?

To create graphs on our COVID-19 data, we utilized Tableau to build visualizations and dashboards.

Sentiment Analysis

We used the Twitter Developer Platform to collect 186,452 original tweets using the rtweet package in R. When we scraped these tweets, we considered tweets that contained keywords associated with COVID-19 vaccines, COVID-19 safety procedures, and lifestyle during the COVID-19 pandemic. We also only considered tweets that were in English. Retweets were disregarded in our data scraping. Tweets collected were from a range of dates between 11/27/2022 and 12/6/2022. During our scraping process, we were unable to scrape the location as one of the columns.

The scraped data was then divided into two categories: vaccinations and lifestyle.

Vaccine	Keywords	Total Collected Tweets	Total Cleaned Tweets *after sentiment analysis
Pfizer	pfizer, Pfizer-BioNTech, BioNTechpfizer	10614	7611
Moderna	Moderna, moderna_tx, Moderna-NIAID, NIAID, NIAID-Moderna	8373	6111
Johnson & Johnson	Johnson & Johnson, Johnson and Johnson, Janssen, Janssen Pharmaceutical, J&J	27798	16829
Oxford-AstraZeneca	OXFORDVACCINE, Oxford-Astraeneca, OxfordAstraZeneca, AstraZeneca, Vaxzevria, Covishield	1738	1145
SputnikV	Sputnik V, sputnikv, sputnikvaccine	61	38
Covaxin	covaxin, BharatBiotech	310	205
Sinovac	coronavac, sinovac	865	65

Tweet Topic	Keywords	Total Collected Tweets	Total Cleaned Tweets *after sentiment analysis
Hygiene	hand sanitizer, sanitizer, wash hands, wash face, soap, soap water, hand soap, sanitize	23258	14139
Wear Mask	mask, wearamask, masking, N95, face cover, face covering, face covered, mouth cover, mouth covering, mouth covered, nose cover, nose covering, nose covered, cover your face, coveryourface	47438	34965
Travel	travel, outing, camping, air-travel	27661	16593
Social Distancing	social distancing, physical distancing, 6 feet, social distance, physical distance	15474	9350
Social Gathering	social gathering, gathering, party, restaurant	24617	16864

Data Processing Procedure – Vaccination Tweets

- **Removing unnecessary columns:** We removed any columns that were unnecessary to our analysis while retaining relevant ones such as the tweet text and the date it was tweeted.
- **Data cleaning:** We extracted the tweet text and converted it to lowercase, removed special characters, Twitter mentions, hashtags, punctuation, numbers, URLs, and unnecessary spaces, replaced “&” symbols with “and”, removed stop words, and lemmatized words.
- **Tokenization:** We used a function in R to unnest each text value into separate words.
- **Sentiment analysis:** We used the afinn sentiment analysis package to obtain the sentiment score for each tweet.

Data Processing Procedure – COVID-19 Data

- **Calculating daily values:** We used functions in Excel to calculate daily values from running total columns.

- **Joining data together:** We joined the data together on the country and date fields, supplementing our original datasets with other data (i.e. country data with population numbers).

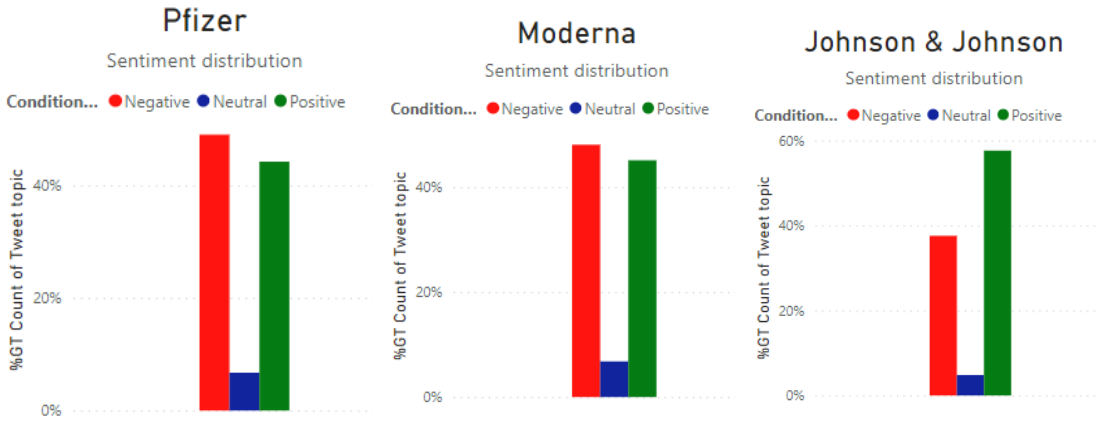
Analysis

Sentiment Distribution for Vaccination Tweets

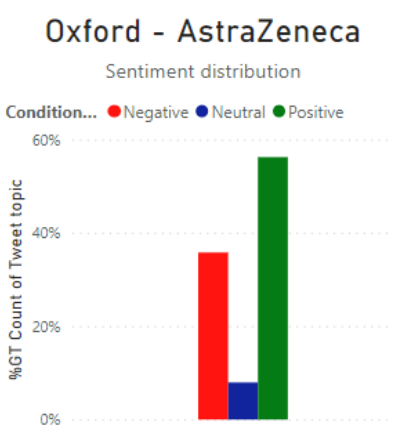
We plotted the distribution of positive, neutral, and negative sentiment across different vaccines. Pfizer and Moderna both had negative sentiment as the highest category, while Johnson & Johnson, Oxford-AstraZeneca, Sputnik V, Covaxin, and Sinovac all had positive sentiment as the highest category.

In comparison to Sattar and Arifuzzaman's analysis where they found the neutral sentiment regarding vaccinations to be the highest, we found the neutral sentiment to be consistently the lowest across all vaccinations. However, Sattar and Arifuzzaman found that positive sentiment outweighed negative sentiment, indicating that majority of people felt positive about the push for vaccination against COVID-19. Given that Johnson & Johnson, Oxford-AstraZeneca, Sputnik V, Covaxin, and Sinovac all had positive sentiment outweigh the negative, it can be inferenced that positive sentiment regarding COVID-19 vaccinations remains positive.

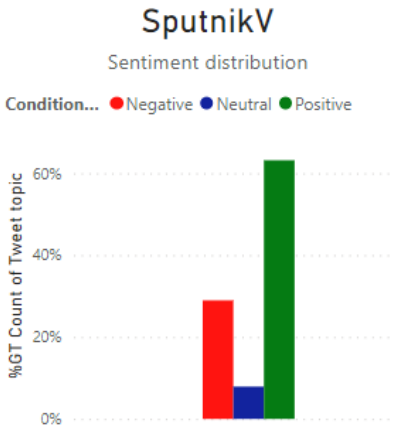
As mentioned previously, we were not able to scrape the location of a tweet. Therefore, we were unable to compare sentiment breakdown by location.



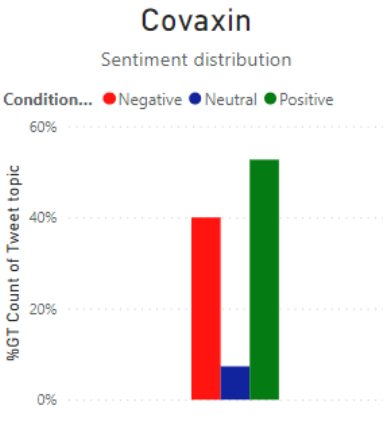
(a) Pfizer *(b) Moderna* *(c) Johnson & Johnson*



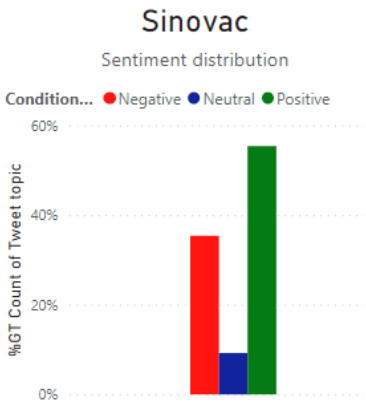
(d) Oxford-AstraZeneca



(e) Sputnik V



(f) Covaxin

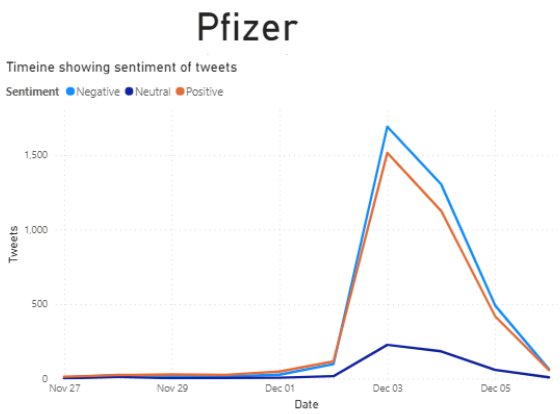


(g) Sinovac

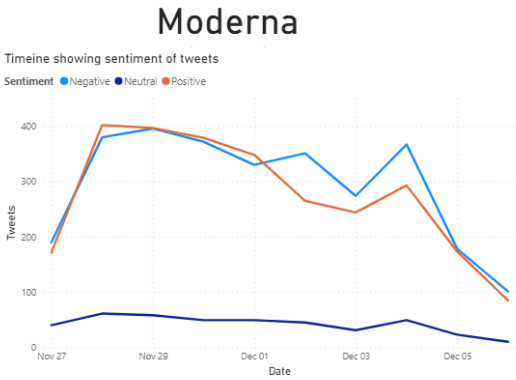
Sentiment Timeline for Vaccination Tweets

We plotted the day-to-day sentiment scores of tweets to analyze the distribution of sentiment across tweets by date. Given that we only worked with a small range of tweet data, the insights from this analysis were not overly significant. However, Pfizer, Johnson & Johnson, and Covaxin displayed a spike in the number of positive and negative tweets in early December. This spike could be attributed to changes in public sentiment regarding COVID-19 vaccines. For example, the FDA just recently updated their authorization for COVID-19 booster shots for kids under 5 years old. Additionally, Pfizer and Moderna have recently been featured in the news regarding Moderna's lawsuit for patent violations by Pfizer.

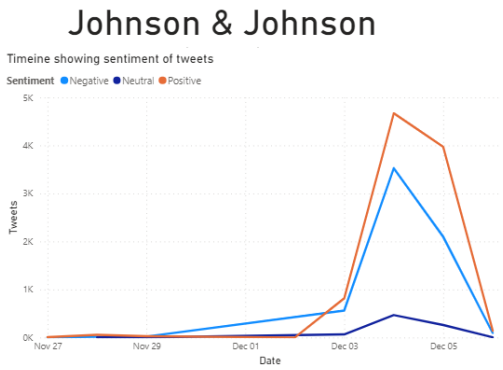
In comparison to Sattar and Arifuzzaman's analysis, we found similar results in relation to the spike in negative or positive tweets regarding vaccines based on corresponding news stories around the time of spikes. However, given that we are looking at a smaller date range than Sattar and Arifuzzaman's analysis, it is difficult to compare results.



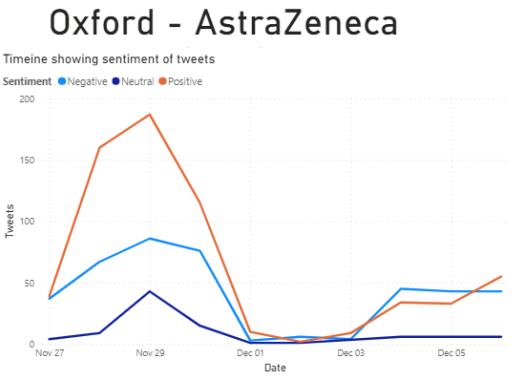
(a) Pfizer



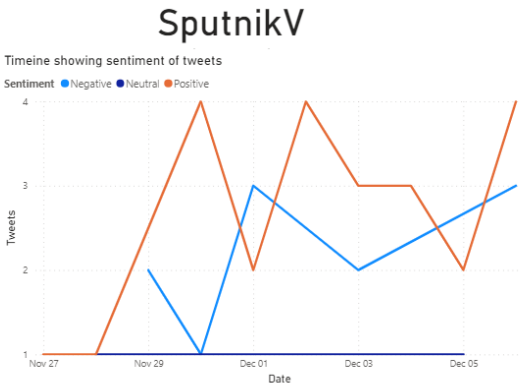
(b) Moderna



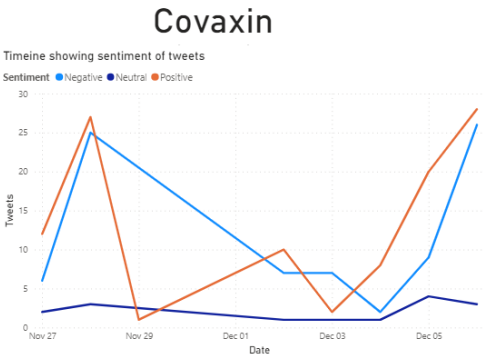
(c) Johnson & Johnson



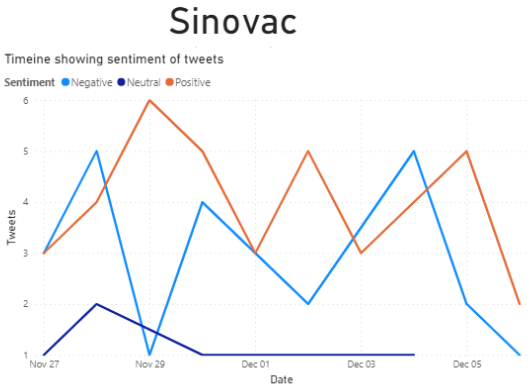
(d) Oxford-AstraZeneca



(e) Sputnik V



(f) Covaxin



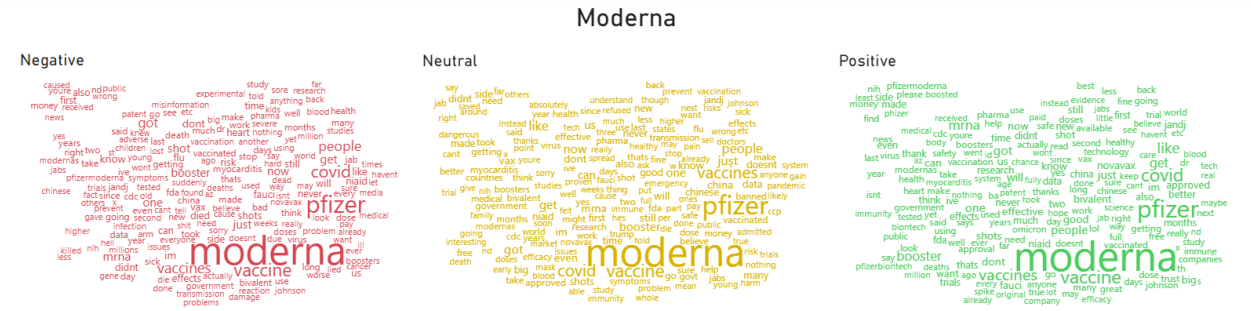
(g) Sinovac

Top Frequency Words (Positive, Negative, Neutral) for Vaccination Tweets

We visualized the high frequency word distribution for each vaccine in a word cloud, categorizing the tweets by negative, neutral, and positive words. The generated word clouds did not produce significant results given that the most tweeted words were related to the vaccines themselves (i.e. “pfizer” was the highest tweeted word regarding the Pfizer vaccination).

In comparison to Sattar and Arifuzzaman’s analysis, we obtained similar results of not produing any significant or useful information from the word clouds. However, Sattar and Arifuzzaman were able to see a pattern of tweets about side effects from vaccinations in the negative word distributions. We were not able to pick out any defining negativity about the vaccinations from any of the negative word distributions. This could be attributed to the difference in when the tweets were scraped. Given that Sattar and Arifuzzaman scraped the data when COVID-19 vaccinations were fairly new, discussion about the side effects would have been more prevalent. By comparison, many people are already vaccinated and discussion on side effects is not as prevalent now.

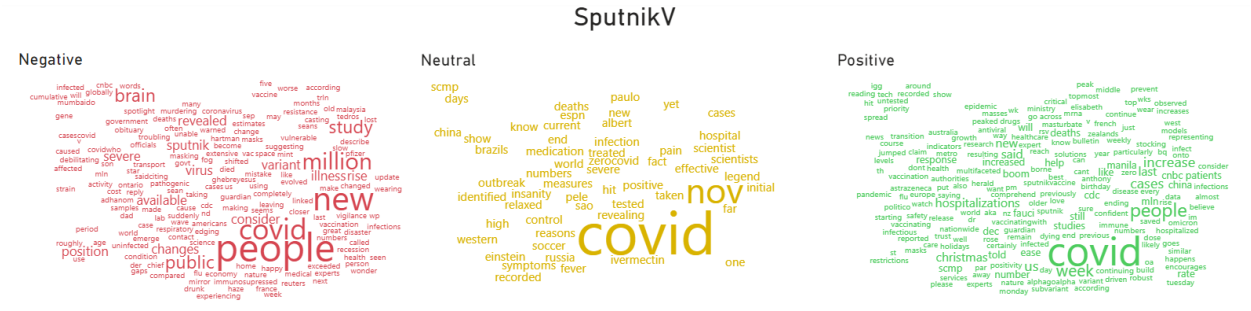
(a) Pfizer



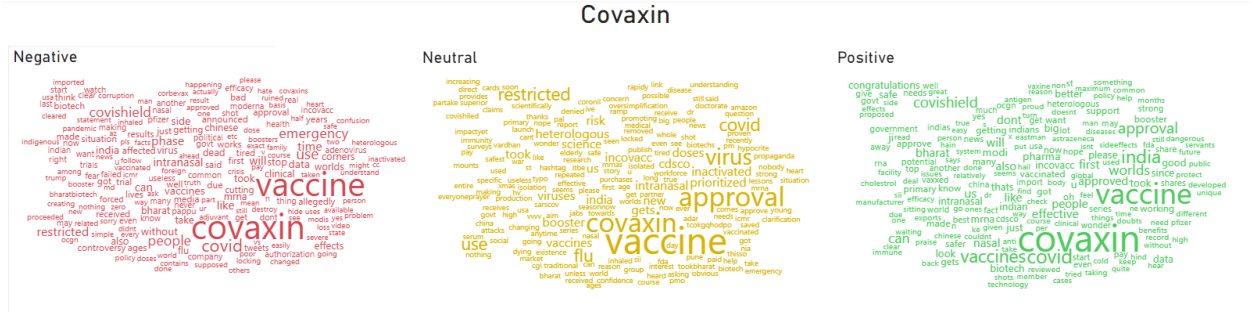
(b) Moderna

(c) Johnson & Johnson

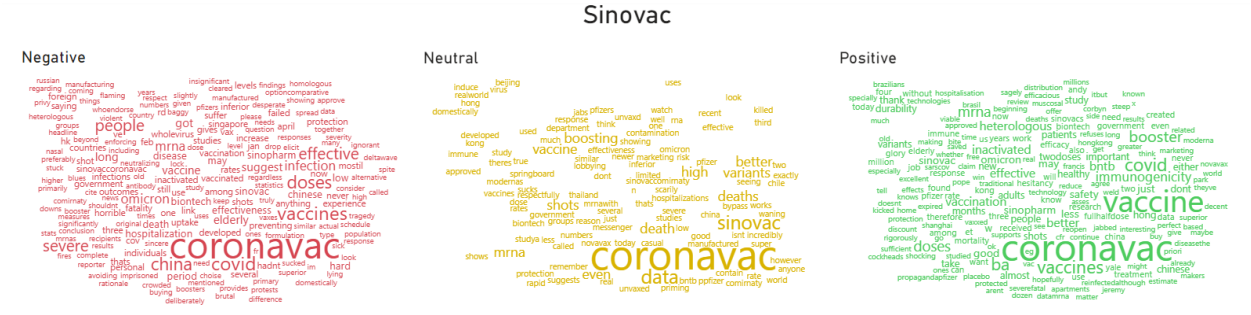
(d) Oxford - AstraZeneca



(e) Sputnik V



(f) Covaxin

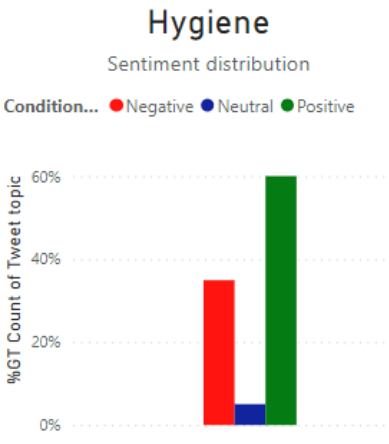


(g) Sinovac

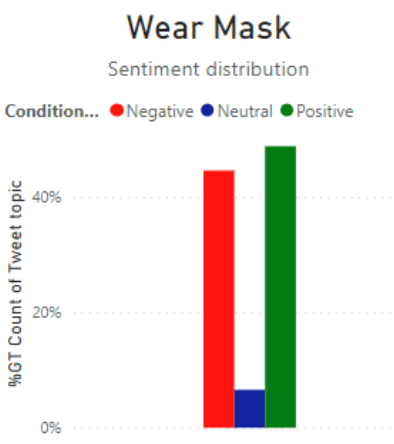
Sentiment Distribution for Post-Vaccination Tweets on Healthy Lifestyle

We plotted the distribution of positive, neutral, and negative sentiments across various post-vaccination and lifestyle-related tweets. Across all categories, Hygiene, Wear Mask, Travel, Social Distancing, and Social Gathering, positive sentiment had the highest percentage of tweets. Most significantly, positive sentiment regarding hygiene, travel, social distancing, and social gathering both made nearly or over 60% of total tweets. This reflects people's favorable outlook on hygiene, traveling, social distancing, and social gatherings. Although wearing a mask also scored high on number of positive sentiment tweets (nearly 50%), negative sentiment also scored high with just over 40% of tweets. The high amount of negative sentiments may be attributed to the unwillingness of people to revert to wearing masks again.

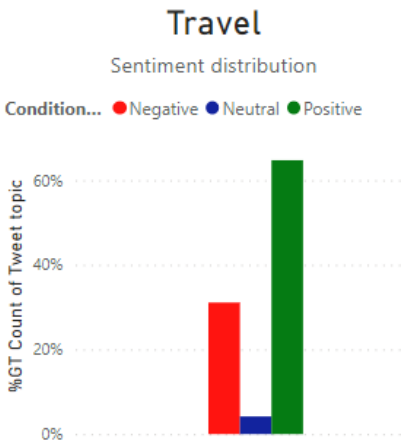
In comparison to Sattar and Arifuzzaman's analysis, some of the results (specifically, social distancing and hygiene) similarly scored high on percentage positive sentiments. However, Sattar and Arifuzzaman's results had a higher proportion of neutral sentiments.



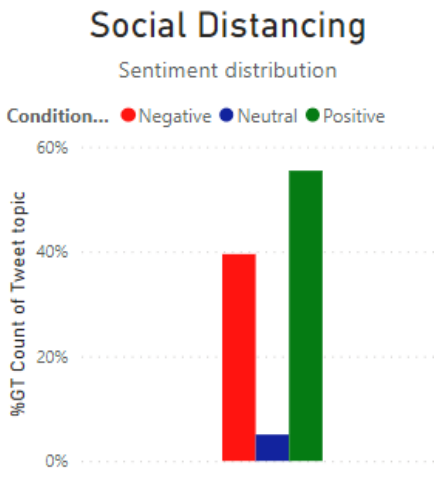
(a) Hygiene



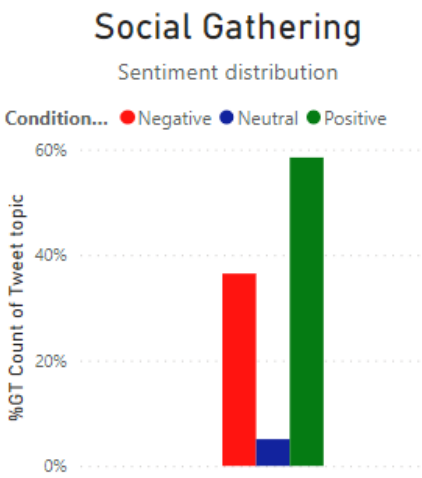
(b) Wear Mask



(c) Travel



(d) Social Distancing

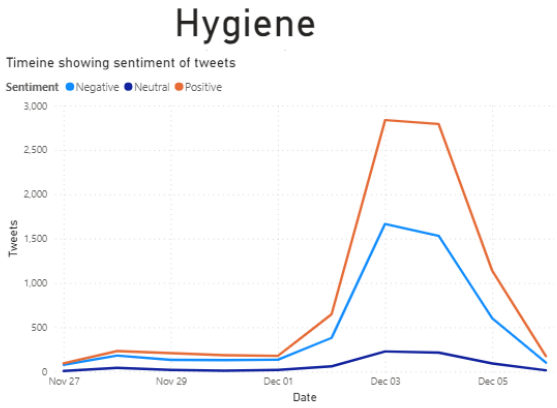


(e) Social Gathering

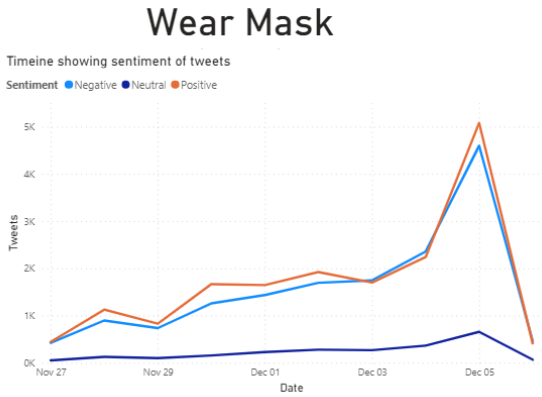
Sentiment Timeline for Lifestyle Tweets

Similar to how we plotted the day-to-day sentiment scores of tweets to analyze vaccination tweets, we also plotted the day-to-day sentiment scores of tweets relating to post-vaccination lifestyle. Again, because we worked with such a small range of tweet data, it was difficult to pull out any relevant insights. One interesting notation is that there was a spike in the number of negative and positive sentiment tweets in the month of December across all keywords. This may be related to the increase of COVID-19 cases and respiratory diseases as the United States enters the winter months. Additionally, in early December, the CDC started recommending people to wear masks again. CBS news claims that 14% of Americans now live in “high” COVID-19 community levels, which is an increase from 5% in the previous week. The increase in negative and positive sentiment tweets could be in response to this.

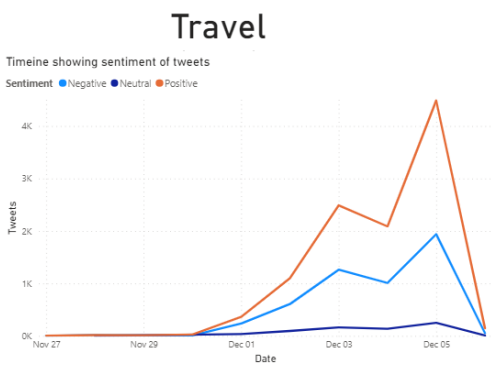
In comparison to Sattar and Arifuzzaman's analysis, our results showed more significant spikes in one period of time in the number of negative or positive sentiment tweets. However, this could be attributed to the smaller time frame of tweets that we scraped for our analysis. If we scraped more tweets across a longer frame of time, our results would most likely mimic Sattar and Arifuzzaman's.



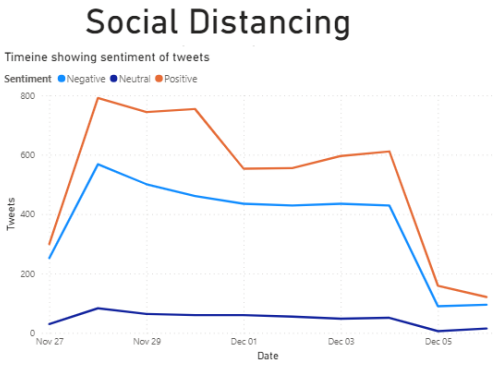
(a) Hygiene



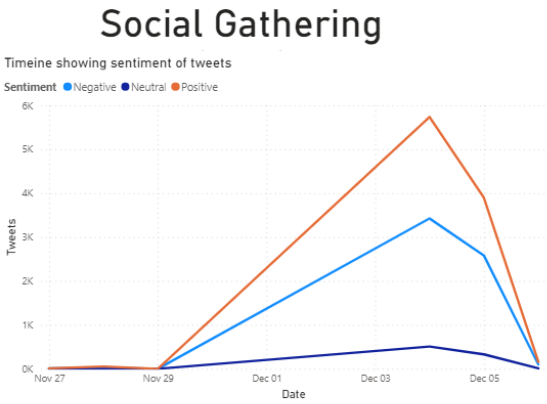
(b) Wear Mask



(c) Travel



(d) Social Distancing

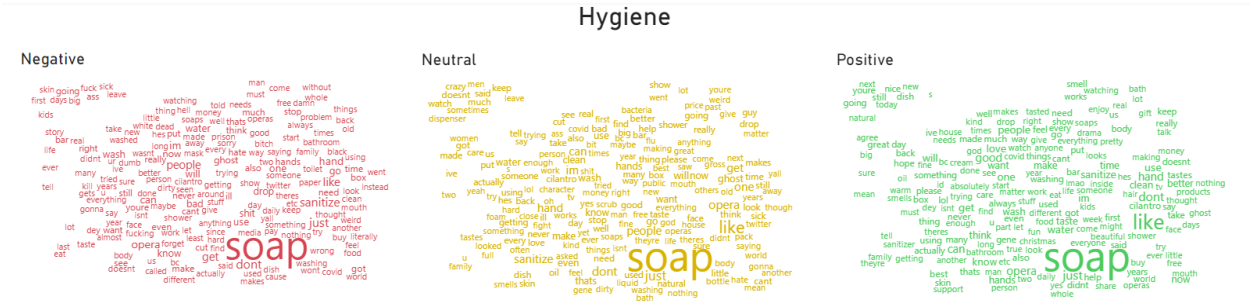


(e) Social Gathering

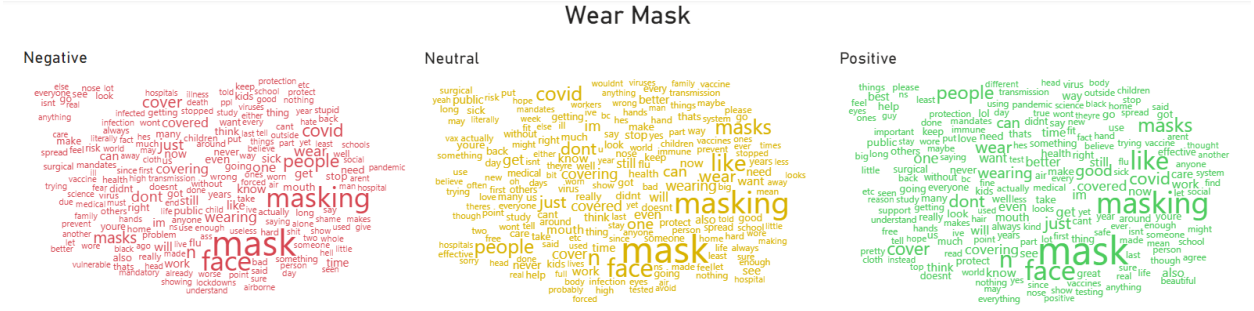
Top Frequency Words (Positive, Negative, Neutral) for Post-Vaccination Tweets on Lifestyle

Similar to the process followed for the top frequency words for tweets relating to vaccinations, we visualized the top frequency words by positive, negative, and neutral sentiment for tweets relating to post-vaccination lifestyle. Again, each topic was categorized by negative, neutral, and positive words. Once again, the generated word clouds did not produce overly significant results. The distribution of words across negative, neutral, and positive sentiments did not produce many discernable differences across word clouds. For example, the travel word cloud all pulled “camping”, “outing”, and “travel” as three of the most tweeted words across all sentiment categories.

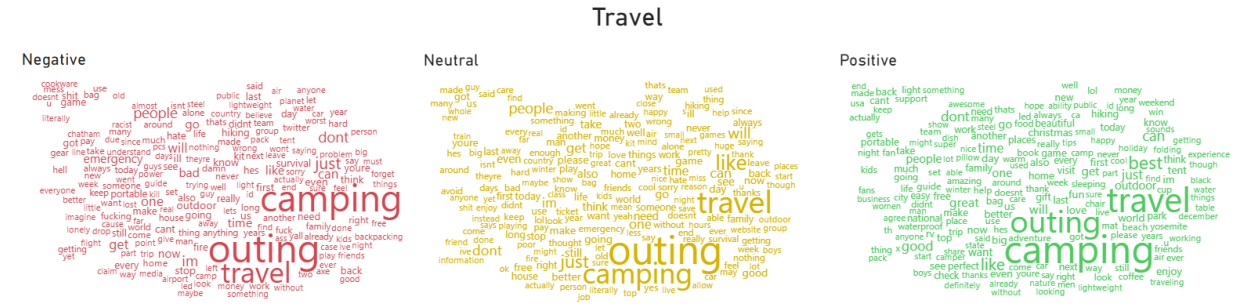
In comparison to Sattar and Arifuzzaman’s results, our results were not as significant. The word clouds created in the study showed strong sentiments regarding death, travel bans, hospitals, etc., while our word clouds showed more generalized language about the topic (i.e. tweets about “hygiene” talked about soap, hands, and washing). The difference in strong sentiments could be because more people during the start of vaccinations were tweeting about the relevant news topics such as travel bans from countries, the number of hospitalizations, etc. Now, most countries do not have travel bans and the number of hospitalizations has been decreasing.



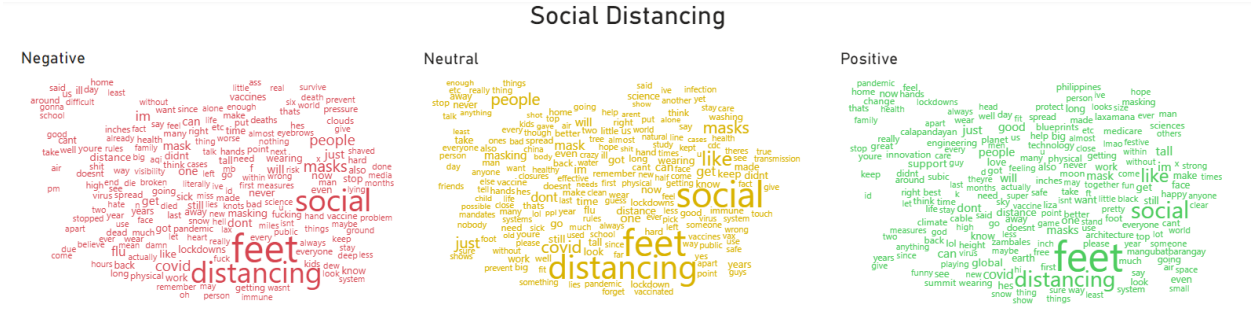
(a) Hygiene



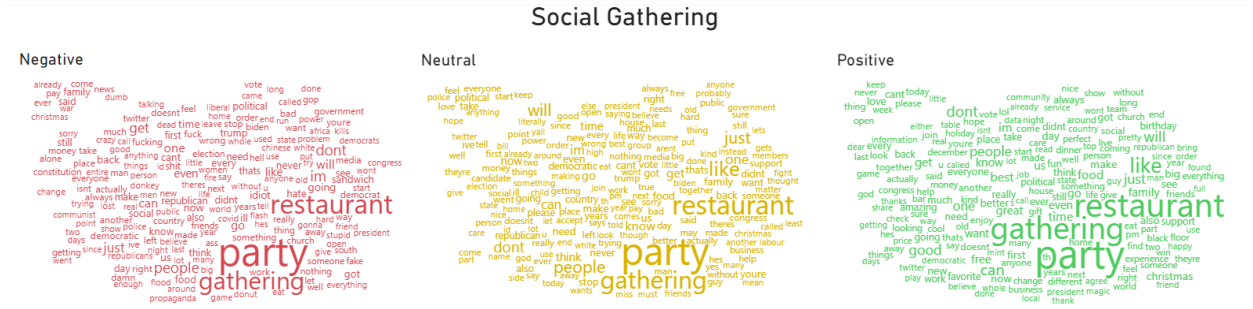
(b) Wear Mask



(c) Travel



(d) Social Distancing

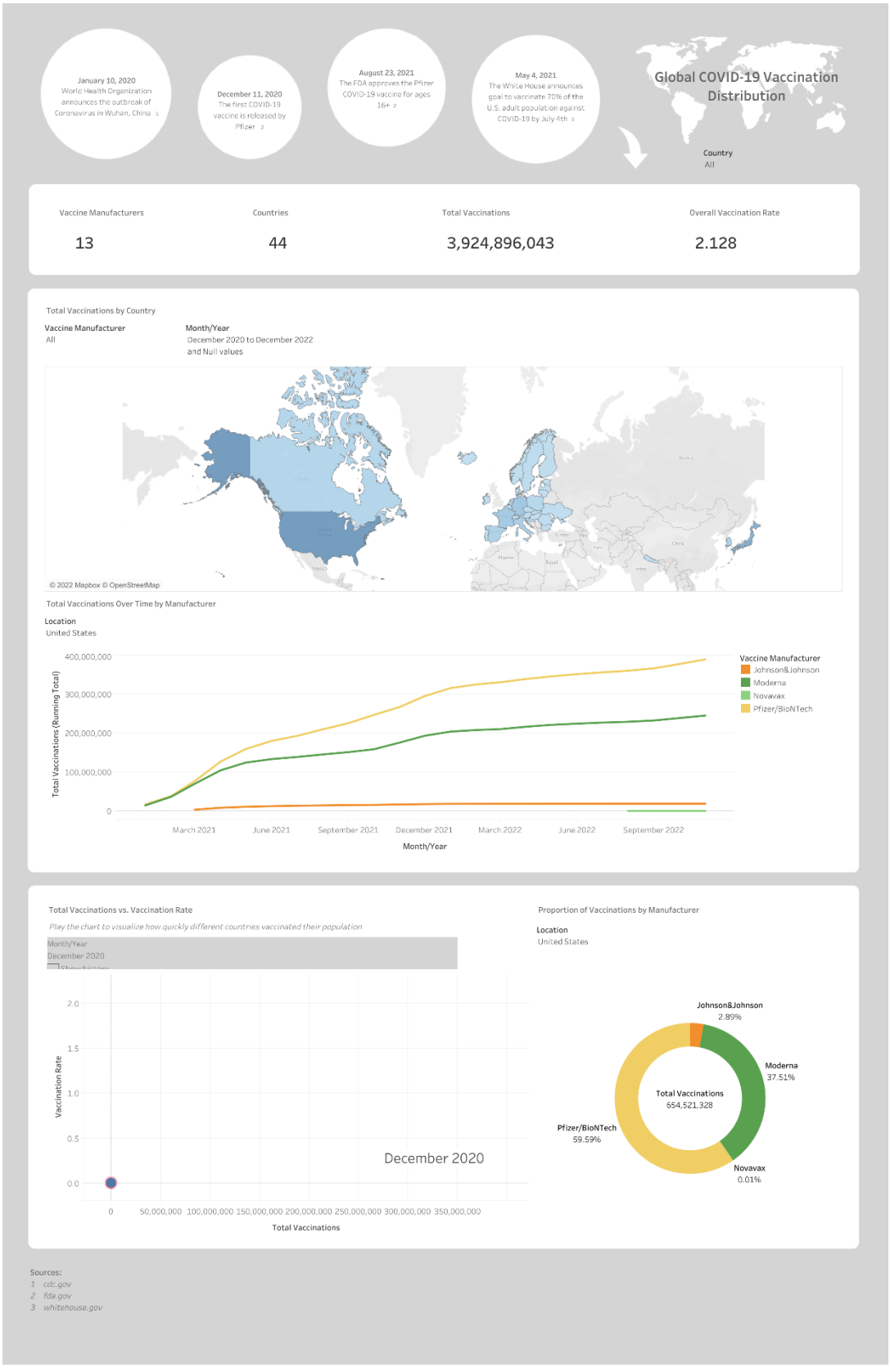


(e) Social Gathering

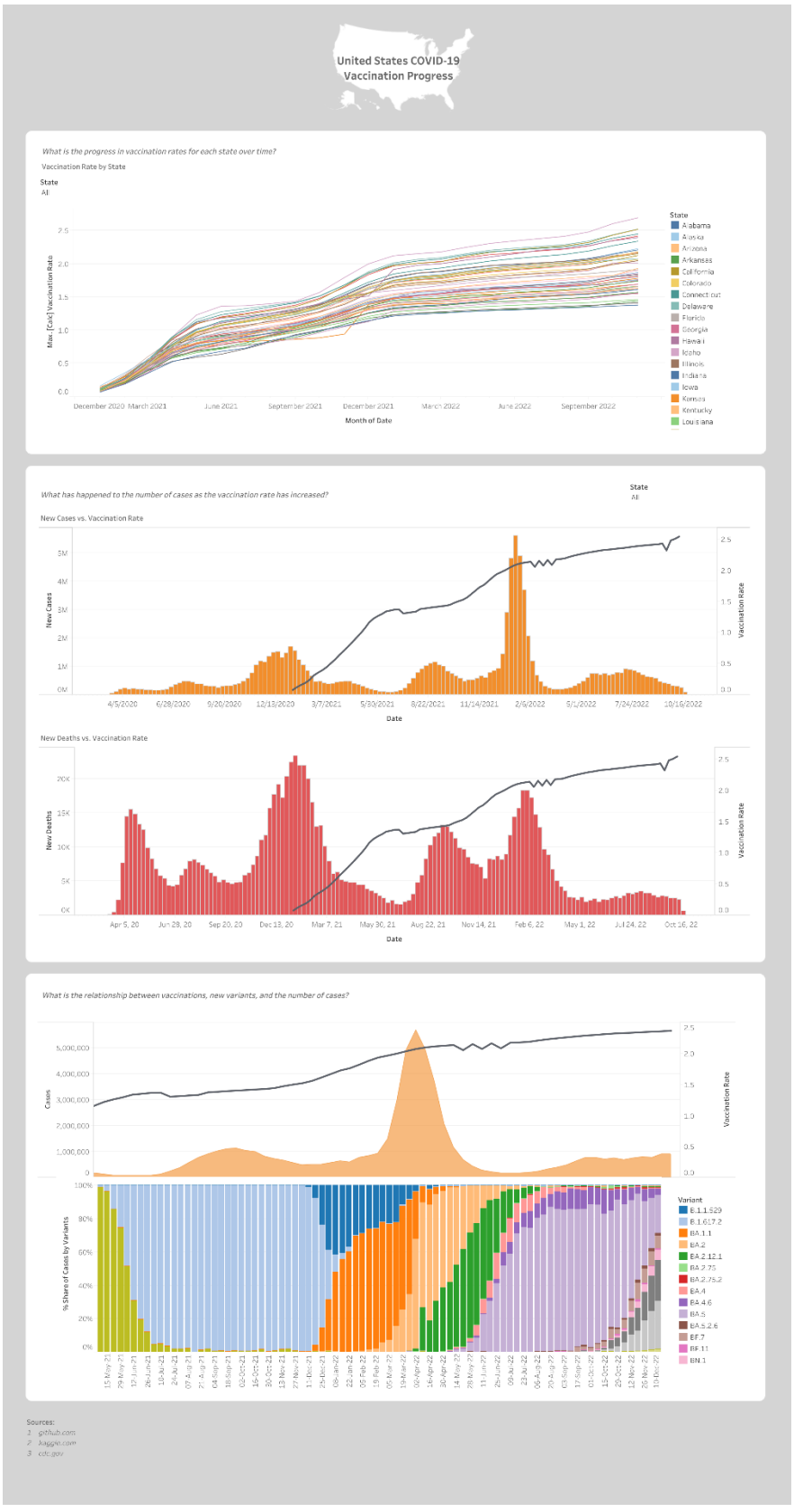
COVID-19 Vaccination Data Dashboards

To visualize our COVID-19 vaccination data, we created three comprehensive dashboards to address our research questions.

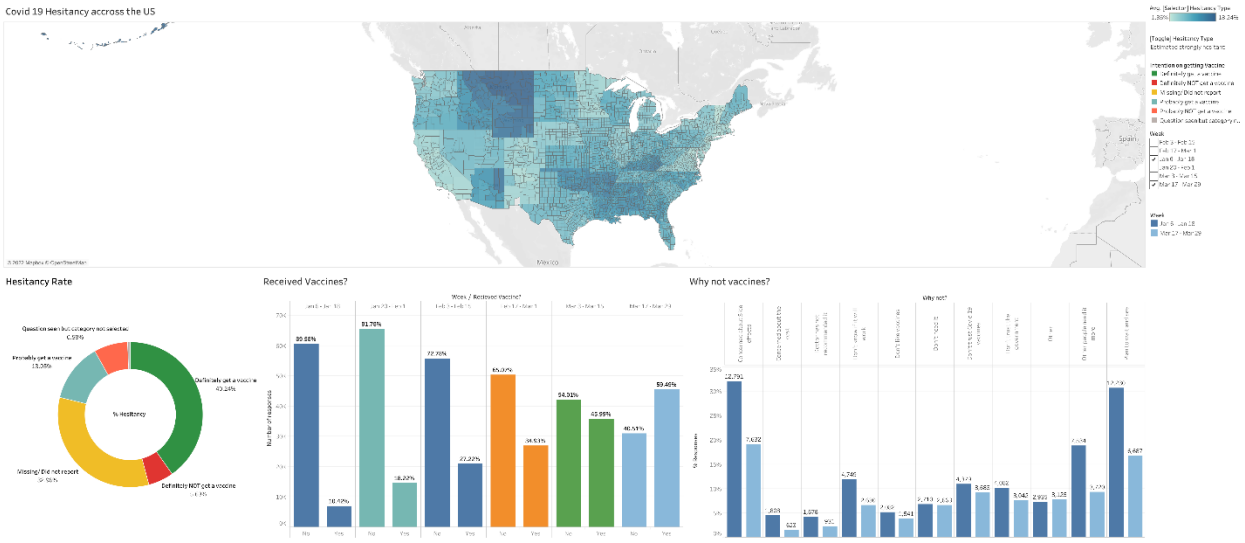
- **Global COVID-19 Vaccination Distribution**
  Link: https://public.tableau.com/app/profile/hannah.wirth/viz/GlobalCOVID-19VaccinationDistribution/GlobalCOVID-19VaccinationDistribution
- **United States COVID-19 Vaccination Progress**
  Link: https://public.tableau.com/app/profile/hannah.wirth/viz/UnitedStatesCOVID-19VaccinationProgress/UnitedStatesCOVID-19VaccinationProgress
- **COVID-19 Vaccine Hesitancy**
  Link: https://public.tableau.com/app/profile/dikshant.joshi/viz/shared/RD983RJX2



(a) Global COVID-19 Vaccination Distribution



(b) United States COVID-19 Vaccination Progress



(c) COVID-19 Vaccination Hesitancy

Summary of Findings and Implications

Overall, our results varied drastically from what Sattar and Arifuzzaman found in their study. Although the results were different, they were not entirely useless. The increases in negative and positive sentiments tweets regarding vaccinations and lifestyle in the month of December line up with recent CDC guidelines for masking in areas with high levels of COVID-19 transmission and changes to vaccination regulations (i.e. who qualifies for vaccinations now). We also attributed our differences in results to the following reasons:

1. We scraped tweets from a different timeline. The study scraped tweets from the height of the pandemic when COVID-19 vaccinations and adjusted lifestyle were still new and consistently being talked about. Our tweets are scraped from November to December 2022, when COVID-19 vaccinations have been available for over a year and the post-COVID-19 lifestyle is common.
2. We scraped tweets in a much smaller time frame. The original study scraped tweets over 5 weeks while we only scraped tweets in a 1-day span. Scraping from such a small time frame may not fully reflect the sentiment surrounding COVID-19 vaccinations and lifestyle. However, if we were to collect tweets for the entire months of November and December, we would be able to find more insights about how the sentiment regarding COVID-19 vaccinations and lifestyle has evolved.

Based on our results, we came to the following conclusions:

- Positive and negative sentiments for almost all keywords across tweets are the most prevalent in our analysis. Neutral sentiment across keywords was not as prevalent.
- The trend of positive and negative sentiment tweets has increased in the month of December 2022. This may be because of the new spike in cases and CDC guidelines.

- In comparison to the study, there are not many repeated words across keywords and sentiments in our analysis. This may be because that we are not seeing as many deaths, side effects, or extreme responses to COVID-19 and vaccinations, or attributes relating to COVID-19 such as discussion about side effects have become common and are no longer discussed.

If we were to revise our methodology behind this project, we would try again to scrape the location of the tweets. Understanding the location behind where the tweets were coming from may provide legitimacy to claims of increased talk about COVID-19 awareness. For instance, a high proportion of positive or negative tweets about lifestyle are coming from New York City, a high COVID-19 transmission area.

Through our analysis of COVID-19 data, we were not able to generate any unique analyses, but we found that our numbers and analysis matched the global and national COVID-19 statistics. This in itself was a success because of our struggles to manipulate the data in a way that made sense. This also indicates that if we were to take our analysis further, our starting base of data and baseline analysis of cases, deaths, vaccination rate, etc. would be correct and allow us to complete more complex analyses.

We were able to conclude the following answers:

1. ***What has the diffusion of vaccines across the world looked like?***

Based on the 44 total countries in our dataset, the United States had the highest number of total vaccinations. We also found that a large portion of Europe has been vaccinated. A notable omission is that the dataset did not include any data on China or India vaccinations.

2. ***What is the progress in vaccination rates for each state over time?***

The vaccination rates in each state has increased over time. Vermont has the highest vaccination rate out of all 50 states.

3. ***What has happened to the number of cases as the vaccination rate has increased?***

As the vaccination rate has increased, the number of cases has declined. One discrepancy in this is a spike in cases in January 2022. Similarly, as the vaccination rate has increased, the number of deaths has declined.

4. ***What is the relationship between vaccinations, new variants, and the number of cases?***

As time goes on from May 2021, we can see that as the number of cases spikes in February 2022, different types of variants emerge. Additionally, as the vaccination rate increases, the number of new types of variants increases. In December 2022, there are more than 10 different variants.

5. ***How has COVID-19 vaccine hesitancy evolved across the US?***

While we were unable to find how hesitancy has evolved given our limited access and availability to data, we were able to determine the top reasons why people were hesitant to get COVID-19 vaccinations and the distribution across different states for vaccine hesitancy.

Lessons Learned

Working with unclean data

One of the biggest issues with this project that we struggled with was working with the unclean data. When scraping and cleaning the Twitter data in R, we ran into issues with the Twitter Developer Platform, data types, and the sentiment analysis. When we started to work with the COVID-19 data for our dashboards, we ran into issues with the data not making sense, containing running totals, and fields not matching up with each other. However, these struggles forced us to think outside the box and come up with unique solutions. For example, when our sentiment analysis seemed incorrect in R, we went back line by line in the code to find the issue. Similarly, to get daily data from a running total, we created new columns with functions to calculate the daily number. We were constantly problem-solving while working with both types of data.

Replicating an analysis

We found replicating the report's analysis a challenge given that we were working with different parameters than Sattar and Arifuzzaman. For instance, their analysis looked at 1.2 million tweets collected over 5 weeks while our analysis used 186,452 tweets. We were constrained by our resources, Twitter scraping limit, and other factors such as time range. It was difficult to compare tweets from the height of the pandemic and the start of COVID-19 vaccinations to the current low-masking and high vaccination percentage. This required us to think about the reasoning behind the discrepancies in results and why they were happening. We had to critically think and conduct external research to be able to interpret our results.

Finding Datasets

After we determined that we couldn't calculate daily cases and deaths from running totals in the original dataset, we sourced a new dataset that contained daily cases and deaths already as separate data points. Ultimately, we were able to find a CDC data source that we joined on our original data. We also had difficulty finding data to visualize COVID-19 vaccine hesitancy. Most of the datasets we found were not available to us or they didn't show enough information about the vaccine hesitancy. After a couple of days of searching for datasets, we were able to find a dataset that showed COVID-19 vaccine sentiment, but not over a span of time. Trying to find data that answered our questions while also modeling the data correctly.

Being Concise with Results

Another challenge we faced was figuring out how to be concise with our results. We particularly ran into this when creating dashboards for our COVID-19 data. We had a lot of variables to work with and we had to figure out which were the most relevant to answer the questions given to us. We also had to critically think about the visualizations we made and determine whether they

were actually showing useful information to the user. It was easy to make visualizations, but it was more difficult to convey the information in a digestible and comprehensive manner.

REFERENCES

https://www.cdc.gov/media/releases/2022/s1209-covid-vaccine.html

https://dailyvoice.com/new-york/nassau/news/covid-19-cdc-now-recommends-indoor-mask-wearing-in-these-9-ny-counties/850864/

https://www.cnbc.com/2022/12/05/cdc-encourages-people-to-wear-masks-to-prevent-spread-of-covid-flu-rsv.html

https://www.cbsnews.com/news/covid-19-cdc-recommends-masks-new-york-city-los-angeles/

APPENDIX

Supplementary Data Sources

Population data

- https://www.kaggle.com/datasets/tanuprabhu/population-by-country-2020
- https://www.kaggle.com/datasets/peretzcohen/2019-census-us-population-data-by-state

Variants data

- https://data.cdc.gov/Laboratory-Surveillance/SARS-CoV-2-Variant-Proportions/jr58-6ysp

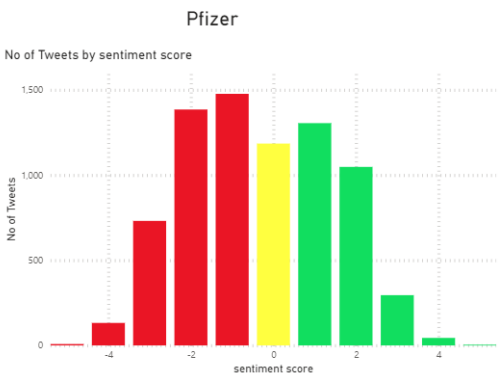
Vaccine hesitancy data

- https://data.cdc.gov/Vaccinations/Vaccine-Hesitancy-for-COVID-19-County-and-local-es/q9mh-h2tw/data

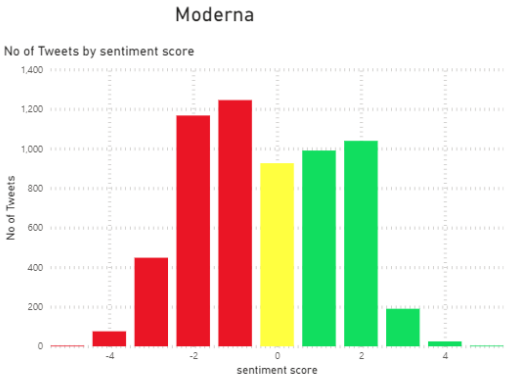
Daily cases and deaths data

- https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36/data

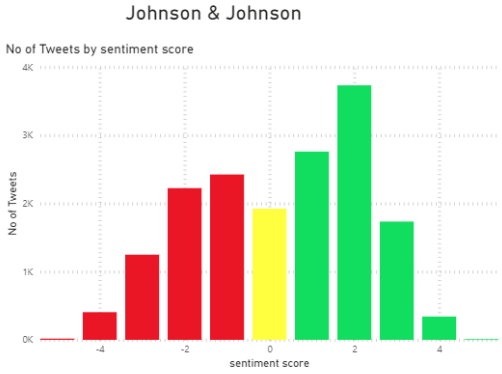
Appendix A: Sentiment Analysis Across Different Vaccines



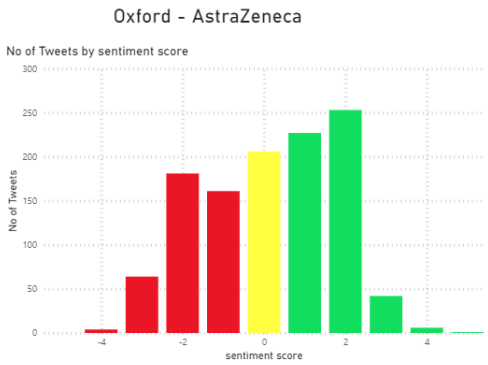
(a) Pfizer



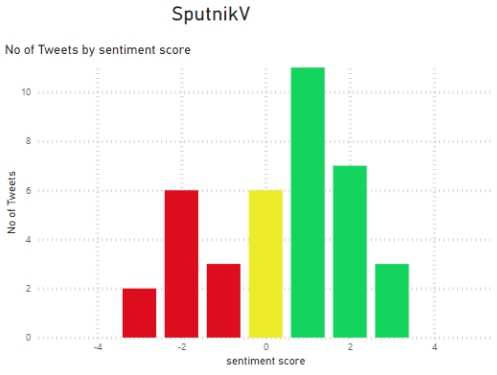
(b) Moderna



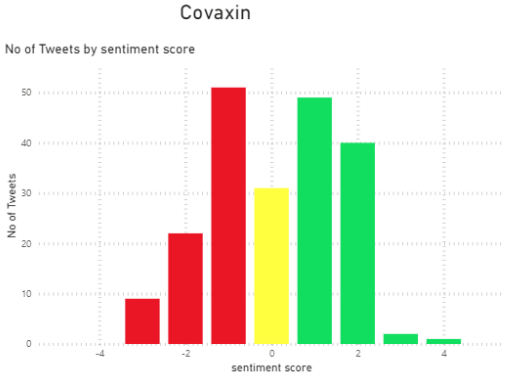
(c) Johnson & Johnson



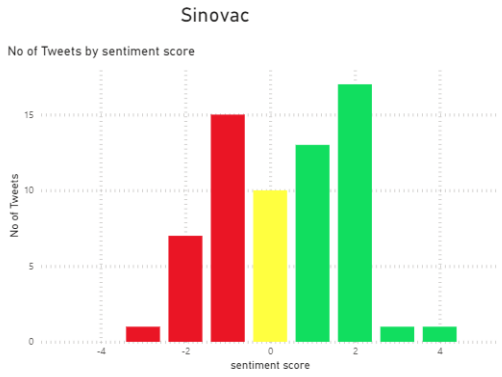
(d) Oxford-AstraZeneca



(e) SputnikV

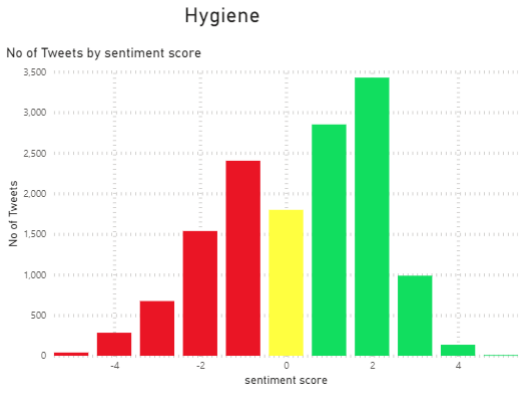


(f) Covaxin

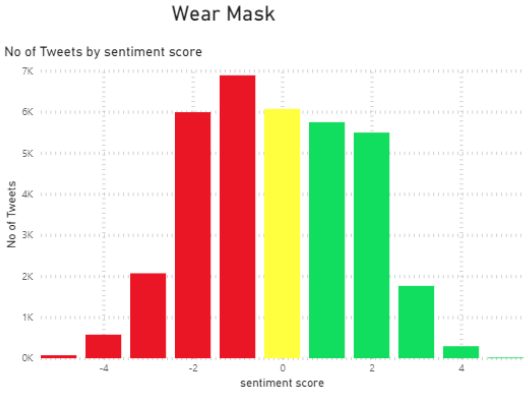


(g) Sinovac

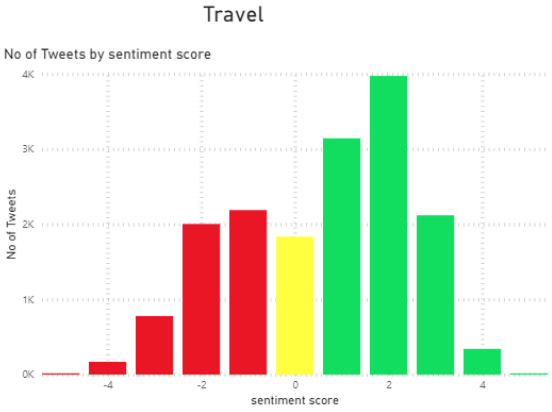
Appendix B: Sentiment Analysis for Healthy Post-Vaccination Lifestyle



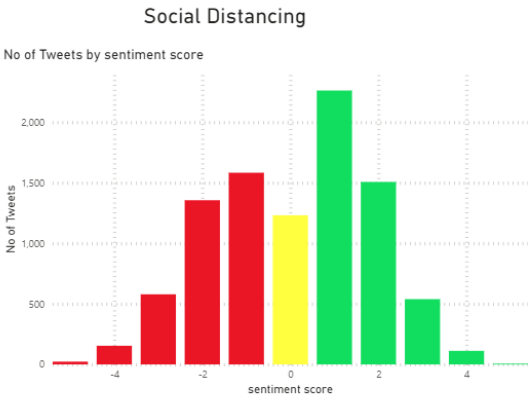
(a) Hygiene



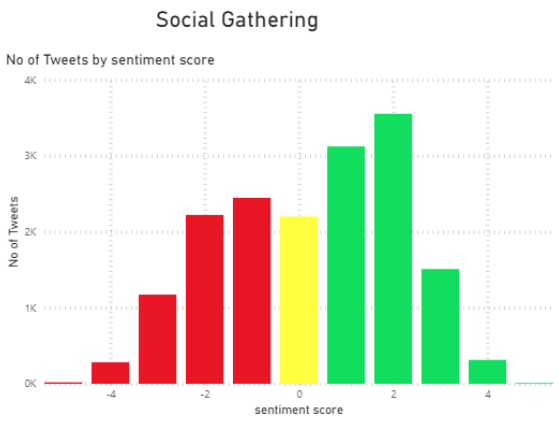
(b) Wear Mask



(c) Travel



(d) Social Distancing



(e) Social Gathering

Code

```
#libraries
library(rtweet)
library(tidyverse)
library(tidytext)
library(tm)
library(SnowballC)
library(wordcloud)
library(RColorBrewer)
library(syuzhet)
library(maps)
library(textstem)
library(webr)
library(systemfonts)
```

```r
#Twitter API access
appname<-""
apikey<-""
apisecret<-""
access_token<-""
access_secret<-""

AuthTwitter <- create_token(
  app = appname,
  consumer_key = apikey,
  consumer_secret= apisecret,
  access_token= access_token,
  access_secret= access_secret)


#Example of function to scrape tweets
hand_soap = search_tweets(q="hand soap",n=30000,include_rts = FALSE,
                          lang ="en", token=AuthTwitter)


#Adding keyword to each tweet row
hand_soap=hand_soap%>%
  mutate(keyword = "hand soap")


#Example of combining rows
total<-rbind(hand_soap,sanitize,mask,wearamask,masking,N95,face_cover,
             face_covering,face_covered,mouth_cover,mouth_covering,mouth_covered,
             nose_cover,nose_covering,nose_covered,cover_your_face)


#Dropping columns
total_cleansed=select(total,-c(entities,metadata,coordinates,place,geo,
                               withheld_in_countries,quoted_status))
```

```r
#Cleaning data
#to lower case
text<-total$text
text<-tolower(text)
#remove non ascii characters
text<-gsub('[^\x20-\x7E]','',text)
#remove RT from retweets, unsure if we have retweets but still ran it
text<-gsub("(RT|via)((?:\\b\\W*@\\w+)+)","",text)
#remove mentions
text<-gsub("@\\w+","", text)
#remove hashtags
text<-str_replace_all(text,"#[a-z,A-Z]*"," ")
#& to and
text<-gsub("&amp","and", text)
#punctuation & digits
text<-gsub("[[:punct:]]", "", text)
text<-gsub("[[:digit:]]","", text)
#urls
text<-gsub("http\\w+"," ",text)
text<-gsub("[ \t]{2,}", " ",text)
text<-gsub("^\\s+|\\s+$", " ", text)
#spaces
text<-str_replace_all(text," "," ")
#clean stop words
text<-removeWords(text, stopwords("english"))
#lemmatize words
text<-lemmatize_words(text, dictionary = lexicon::hash_lemmas)
#place tweet text back
total_cleansed<-bind_cols(total_cleansed, text, id=NULL)
#rename column
total_cleansed<-total_cleansed %>% rename(text_cleansed=...38)
```