BIS634 Final Project

COVID-19 Vaccine Tweets Analysis

Shurui Wang (sw2349)

December 15, 2022

1 Introduction

This research is mainly based on the COVID-19 vaccine tweets dataset. After cleaning and integrating the data, I used Plotly to visualize the data and created several interactive plots. To better present the results, I have created a local Web application with Python Flask that contains everything I have obtained from this research.

2 Data

In this research, I used a Kaggle dataset named COVID-19 vaccine tweets. This dataset contains tweets about the COVID-19 vaccines used in the entire world, with the following vaccines:

- Pfizer/BioNTech
- Sinopharm
- Sinovac
- Moderna
- Oxford/AstraZeneca
- Covaxin
- Sputnik V

2.1 Basic Information

2.1.1 Data Description

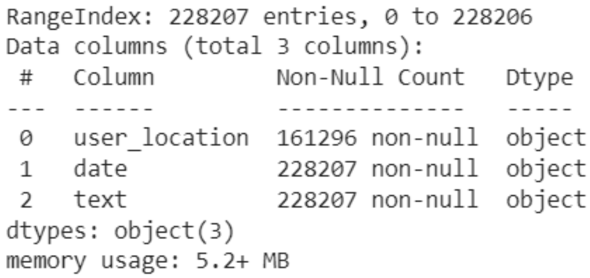
The original dataset has 228207 rows and 16 columns. Each row represents a tweet, with user information, date(from 2020-12-12 to 2021-11-23), and the tweet's attributes recorded in different columns. In this project, I only need three columns out of 16 columns, which are text, user_location, and date.

text: The complete text of a tweet. Hashtags, emojis and symbols are included.

date: The exact date and time of the tweet.

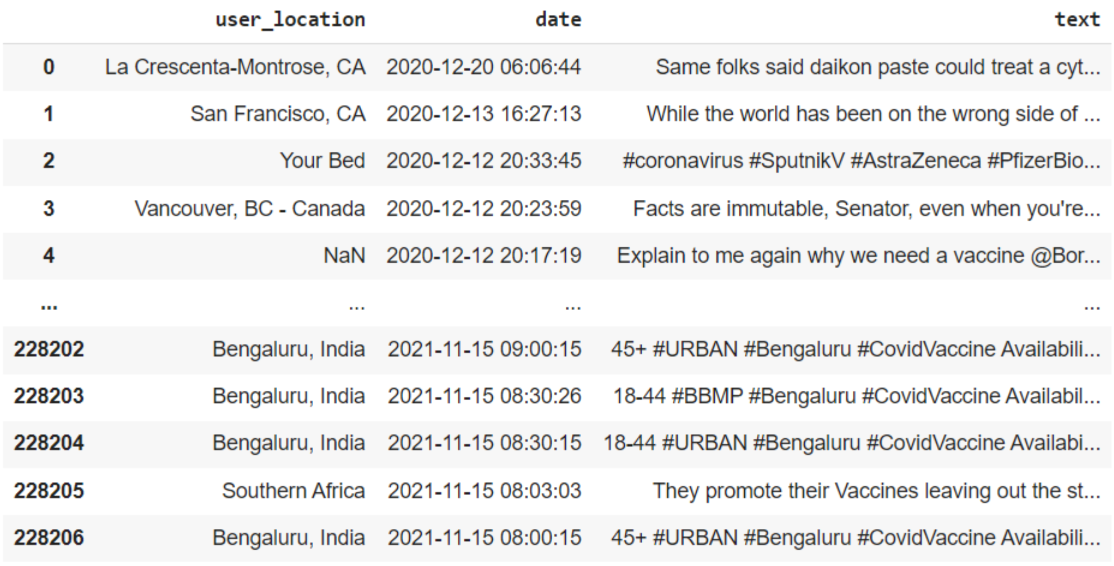
user_location: The location user selected when tweeting, which is not necessarily the user's actual location.

Here is an overview of the dataset:



```
RangeIndex: 228207 entries, 0 to 228206
Data columns (total 3 columns):
 #   Column         Non-Null Count   Dtype 
---  ------         --------------   ----- 
 0   user_location  161296 non-null  object
 1   date           228207 non-null  object
 2   text           228207 non-null  object
dtypes: object(3)
memory usage: 5.2+ MB
```

Figure 1: Dataset Information



	user_location	date	text
0	La Crescenta-Montrose, CA	2020-12-20 06:06:44	Same folks said daikon paste could treat a cyt...
1	San Francisco, CA	2020-12-13 16:27:13	While the world has been on the wrong side of ...
2	Your Bed	2020-12-12 20:33:45	#coronavirus #SputnikV #AstraZeneca #PfizerBio...
3	Vancouver, BC - Canada	2020-12-12 20:23:59	Facts are immutable, Senator, even when you're...
4	NaN	2020-12-12 20:17:19	Explain to me again why we need a vaccine @Bor...
...	...	...	...
228202	Bengaluru, India	2021-11-15 09:00:15	45+ #URBAN #Bengaluru #CovidVaccine Availabili...
228203	Bengaluru, India	2021-11-15 08:30:26	18-44 #BBMP #Bengaluru #CovidVaccine Availabil...
228204	Bengaluru, India	2021-11-15 08:30:15	18-44 #URBAN #Bengaluru #CovidVaccine Availabi...
228205	Southern Africa	2021-11-15 08:03:03	They promote their Vaccines leaving out the st...
228206	Bengaluru, India	2021-11-15 08:00:15	45+ #URBAN #Bengaluru #CovidVaccine Availabili...

Figure 2: Vaccine Tweets Dataset Overview

Noticed that there are missing values in the user_location column. This is because not all users will include their locations while tweeting. Since the text and date columns don't have missing values, I will leave the user_location column untouched.

2.1.2 Why is it interesting?

The COVID-19 pandemic keeps affecting our lives in the past few years. Vaccinations are essential to protect ourselves from COVID, however, some people may hesitate of taking the vaccine, while others may suffer from adverse effects after taking the vaccine. As shown in the below figure, the vaccination rate gets higher and higher in 2021, how people think about vaccines and their feedback after taking the vaccines are worth looking at. It is also interesting to analyze the variations in the sentiment with respect to time, vaccine, and country.

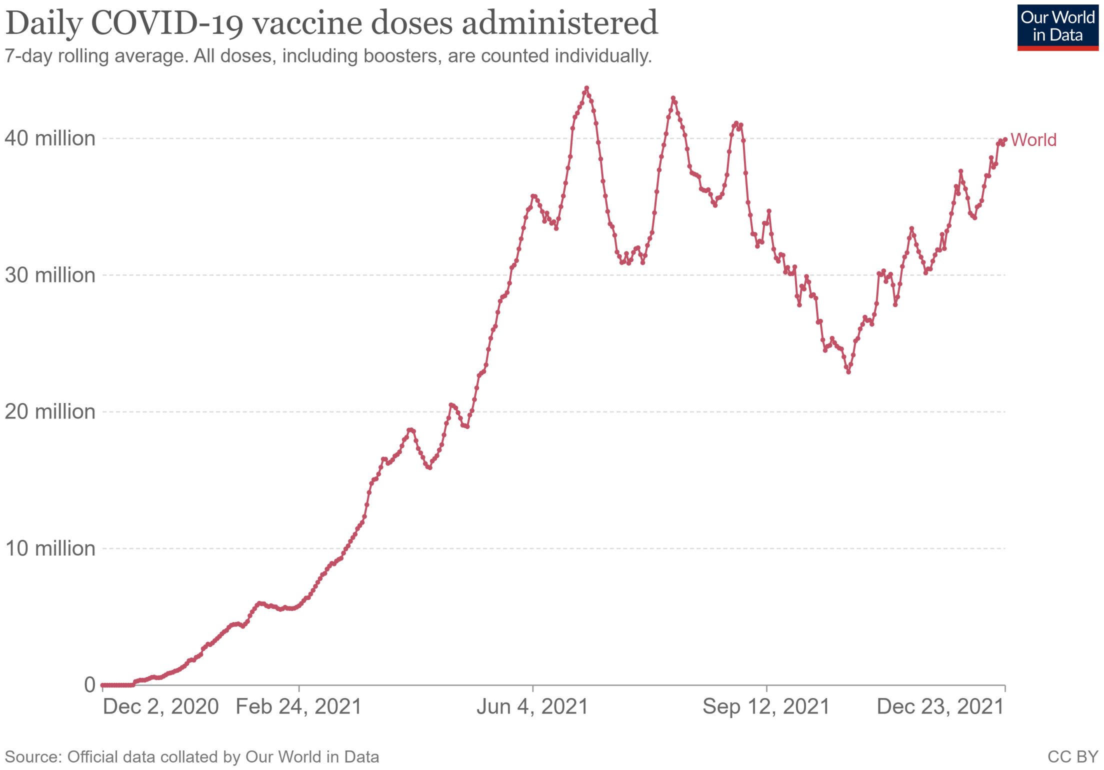


Figure 3: World's Vaccination Rate in 2021

2.1.3 How I acquired the data?

I downloaded this dataset directly from Kaggle(https://www.kaggle.com/datasets/gpreda/all-covid19-vaccines-tweets). The dataset is in .csv format.

2.1.4 Data Collection

The data is collected from Twitter API using tweepy Python package. For each of the vaccine brands, the data provider used relevant search terms (most frequently used in Twitter to refer to the respective vaccine) to acquire data.

2.1.5 The FAIRness of the data provider

The dataset is following the FAIRness principle. The data was well-annotated with metadata and has a clear license. This is a public dataset on Kaggle, with CC0: Public Domain. I can easily access the dataset via Kaggle, without applying for permissions. There's no restriction. There are no certain types of analyses people cannot do with this dataset.

2.1.6 Data Cleaning and Preprocessing

Here are the data cleaning and preprocessing steps before analyzing the data.

- Remove missing/duplicating data.
  Missing data and duplicated data are removed. After removing, there are 116054 rows remaining.
- Standardize the date column into 'yyyy-mm-dd' format.
- Remove URLs, hashtags, emojis, symbols, and punctuations in the text column, and change each row to lowercase.
  This is essential before tokenizing. Otherwise, it will cause trouble in future analyses.
- Tokenize the text column and remove stopwords.
  This step is needed for both building the model and analyzing the results.
- Get sentiment labels for all tweets using the classification model.
  Since the original data don't have labels, classification is needed to analyze the data. All tweets are labeled as -1(negative), 0(neutral), or 1(positive). The details about the model will be discussed in the next section.

2.2 Exploratory Data Analysis

After the data is labeled by the classification model, I did some exploratory data analysis.

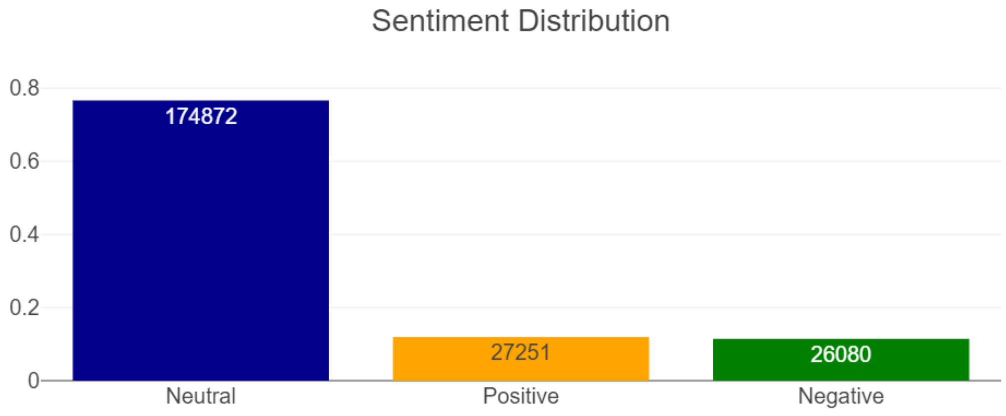


Figure 4: Sentiment Distribution after Labeling

About 75% of the tweets are neutral. The positive tweets are slightly more than the negative tweets, while their proportions are about 12%.

Next, the overall sentiment of each vaccine brand is plotted. If the tweet in the text column contains one of the keywords ('Covaxin', 'Sinopharm', 'Sinovac', 'Moderna', 'Pfizer', 'BioNTech', 'Oxford', 'AstraZeneca', 'Sputnik') it will be counted towards the corresponding vaccine. One tweet might fall into multiple vaccine brands. After generating the subsets for each vaccine, an average sentiment score is calculated. The figure below shows the average sentiment score of each vaccine brand.

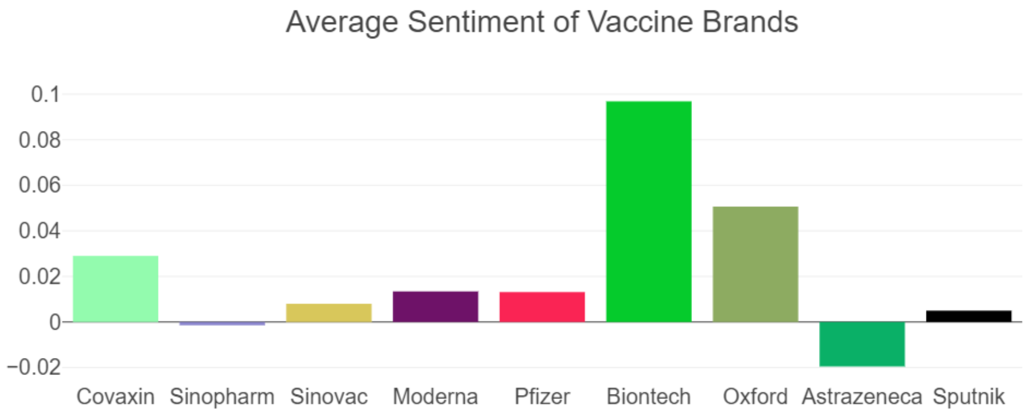


Figure 5: Average Sentiment of Vaccine Brands

From the above plot, BioNTech, Covaxin, and Oxford have the highest average sentiment score, while AstraZeneca has the lowest average sentiment score. AstraZeneca is the only vaccine with a negative score. It is interesting that the same vaccine with different names(Pfizer&BioNTech, Oxford&AstraZeneca) have different average scores. While BioNTech has a high positive sentiment score, Pfizer stays slightly above neutral. Oxford and AstraZeneca have polarized scores.

Similarly, the overall sentiment of each country is plotted. If the tweet in the user_location column contains one of the keywords ('India', 'USA', 'Canada', 'Germany, 'Spain', 'Pakistan', 'UK', 'Brazil', 'Russia', 'Italy', 'Australia', 'France', 'Argentina', 'UAE', 'Israel', 'Mexico', 'Japan') it will be counted towards the corresponding country. One tweet will only be categorized into no more than one country. Here are two special cases: state abbreviations and city names are used to identify the USA, while "UK" and "England" are both used to identify the UK. After generating the subsets for each country, an average sentiment score is calculated. The figure below shows the average sentiment score of each country.

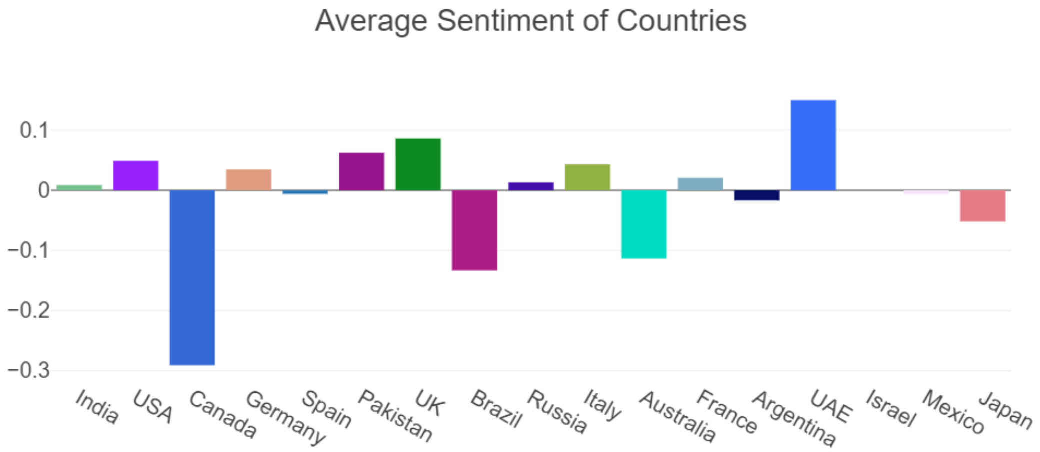


Figure 6: Average Sentiment of Countries

Most of the countries have a positive or negatively neutral overall result. UAE has the highest average sentiment score, and it is the only country with a score higher than 0.1. However, Canada, Brazil, and Australia have a nonnegligible low overall average sentiment score. It may be due to the negative attitude toward vaccination, or it may also be due to the lack of positive/neutral data.

Note that the results might be misleading since the sample size for each vaccine brand/country varies. Some low/high scores may be due to a lack of data. Also since Twitter is not the only social media platform for people to leave feedback, the results cannot represent the overall reputation of a specific vaccine in a specific country.

3 Model

3.1 Basic Information

3.1.1 Training and Validation Data

Since the vaccine tweets dataset is not labeled, I used another dataset to train and validate my model. This dataset is called Complete Tweet Sentiment Extraction Data. I downloaded the dataset directly from Kaggle(https://www.kaggle.com/datasets/maxjon/complete-tweet-sentiment-extraction-data). The dataset is in .csv format and follows the FAIRness principle. The training set contains 90% of the data, while the validation set holds the rest 10%.

3.1.2 Data Cleaning and Preprocessing

This dataset has several columns, however, I only need the ”text” column and the ”new_sentiment” column to train the model.

text: Contains 31329 raw tweets.

new_sentiment: Contains the corresponding sentiment labels for 31329 tweets.

The data cleaning and preprocessing procedures are exactly the same as the procedures for the vaccine tweets dataset. Please refer to the data cleaning subsection in the previous section for more information.

3.1.3 Training and Fine-tuning

After data cleaning and preprocessing, it's time to train the model. I started with a pre-trained model called BERT. BERT stands for "Bidirectional Encoder Representation with Transformers", and it is a pre-trained language model. To put it in simple words, BERT extracts patterns or representations from the data or word embeddings by passing them through an encoder. BERT falls into a self-supervised model. That means, it can generate inputs and labels from the raw corpus without being explicitly programmed by humans. I think BERT is a good fit for predicting sentiments.

My model is built from the BERT model using PyTorch. I adjusted the model parameters and fine-tuned them with the training and validation dataset.

3.2 Model Performance

The model has a 75% training Accuracy, with a training loss of 0.59. As for the validation accuracy, it reaches 80% with a validation loss of 0.48.

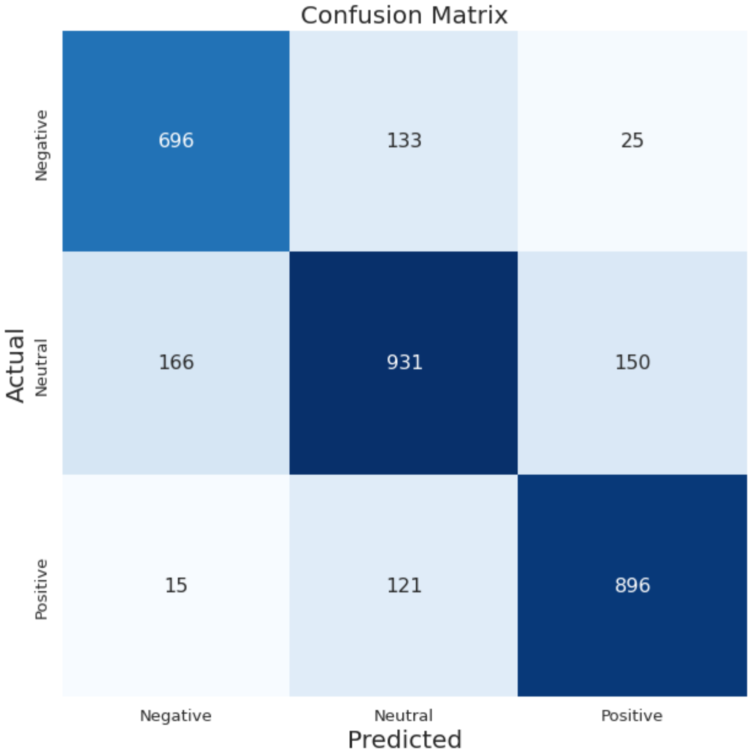


Figure 7: Confusion Matrix on the Validation Dataset

By checking the confusion matrix, I can tell the model can clearly distinguish between positive and negative tweets with very few errors in the bottom left corner and the upper right corner. However, it makes mistakes when distinguishing between negative and neutral, or positive and neutral. This is acceptable since if you check the probability breakdown, the probability for negative&neutral or positive&neutral are very close, and by the nature, the model chose to label the text with the sentiment whose probability is the highest. Also, there are some cases hard to label even for humans. I think the model performance is overall acceptable and can be used to label the COVID-19 vaccine tweets dataset.

4 Data Analysis, Results and Interesting Findings

With the data visualization methods, I analyzed and investigated the dataset, summarized its main characteristics, and compared the changes over time for different categories. I've conducted two analyses, to answer the corresponding research question.

4.1 Word Cloud Analysis

4.1.1 Research Questions

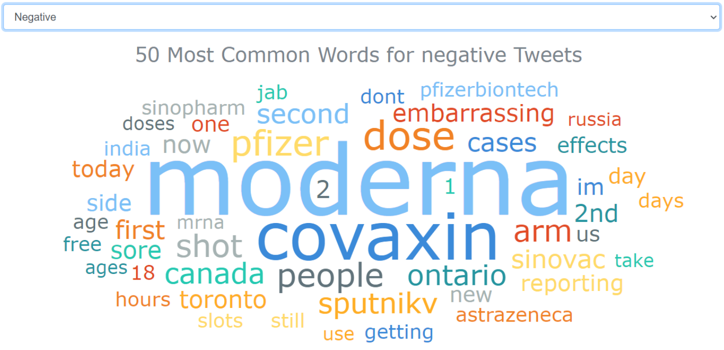
What are the most common words in tweets with different sentiment?

What are the most common words in tweets about different vaccine brands?

What are the most common words in tweets in different countries?

4.1.2 Results and Interesting Findings

I have implemented three interactive word clouds in my Web application. The user can choose from different sentiments, countries, and vaccine brands to generate different word clouds for the top 50 most common words. Word frequencies and percentages of the total are available if the user moves their mouse to the corresponding word. The larger the word is, the higher its frequency is among all qualifying tweets.



Negative

50 Most Common Words for negative Tweets

Figure 8: Word Clouds for Tweets in Japan

1. ”Canada”, ”Ontario” and ”Toronto” appeared on the word cloud for negative tweets. This matched the EDA result that tweets in Canada have the lowest average sentiment score.

2. It seems that all brands of vaccines have positive feedback. ”good” and ”effective” appeared on all word clouds for different vaccine brands.

3. It’s interesting to point out the Japan is the only country whose word cloud has nothing to do with itself. Everything shown on Japan’s word cloud is relevant to the US (e.g. US politicians’ names).

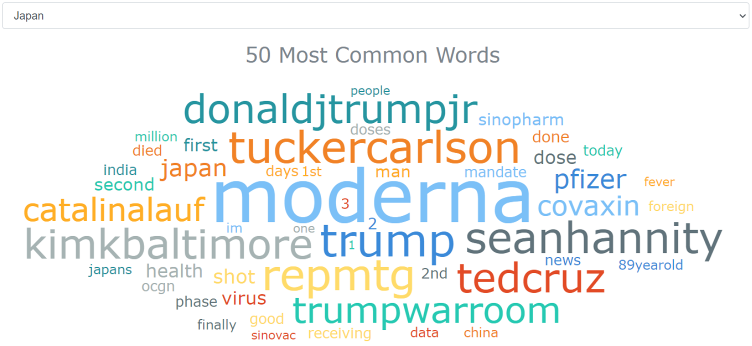


Figure 9: Word Clouds for Tweets in Japan

4. The main topics in the COVID-19 vaccines tweets are availabilities, effectiveness, number of doses, policies, and side effects. Different vaccine brands are also discussed frequently.

4.2 Timeseries Analysis

4.2.1 Research Questions

How does the sentiment change over time?

Do they have a common trend?

4.2.2 Results and Interesting Findings

I have implemented three interactive time series plots in my Web application. The user can either plot the average sentiment over time for different vaccines and different countries, or plot the number of tweets over time in a selected sentiment. Average sentiment scores and tweet counts are available if the user moves their mouse to the corresponding date.

1. As a common trend for Pfizer/BioNTech and Moderna, we can see that in the initial stages the sentiment seems to be above neutral. However, when it comes to the end of the year, the overall sentiment tends to move negatively.

2. As for all other vaccines, the overall sentiment stays neutral. However, this may be due to the lack of data.

3. The countries with an appropriate data size are UK, US, and Canada. They have a common trend where the overall sentiment score moves from the positive side to the negative over time. Canada’s scores are on the negative side all the year. As for other countries, the data size is not enough thus useful results cannot be generated from the plots.

4. The discussion of COVID-19 vaccines on Twitter kept increasing and reached its peak in the middle of 2021, then decreases till the end of the year. When CDC releases new changes in vaccination policies, people are more likely to discuss more on Twitter. Neutral is the dominant sentiment, but it could be argued that Positive Negative sentiments were fairly equal.

5 Server API and the Web Front-end

The local Web application is built with Python flask, along with some HTML and CSS, and some Javascript/JQuery. There are 7 app routes and 7 APIs.

5.1 API

The APIs are used to generate data for interactive plots, or to generate classification output using the model. All outputs are in JSON format.

- @app.route(’/wordcloud_api/<string:vax_name>’): Returns JSON-encoded data generating a word cloud given a vaccine brand.

- @app.route(’/wordcloud_api/senti/<string:senti>’): Returns JSON-encoded data generating a word cloud given a sentiment.

- @app.route(’/wordcloud_api/country/<string:country>’): Returns JSON-encoded data generating a word cloud given a country name.

- @app.route(’/timeseries_api/varianceVStime/<string:vax_name>’): Returns JSON-encoded data generating a timeseries graph given a vaccine brand.

- @app.route(’/timeseries_api/varianceVStime_Country/<string:country>’): Returns JSON-encoded data generating a timeseries graph given a country name.

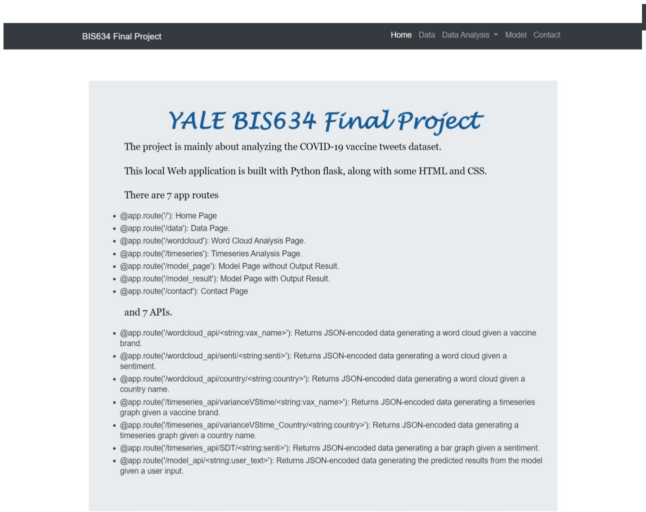
- @app.route(’/timeseries_api/SDT/<string:senti>’): Returns JSON-encoded data generating a bar graph given a sentiment.

- @app.route(’/model_api/<string:user_text>’): Returns JSON-encoded data generating the predicted results from the model given a user input.

5.2 Web Front-end

There are 7 total pages.

- @app.route('/'): Home Page. Displays the structure of the Web application.
- @app.route('/data'): Data Page. Displays everything about the dataset.



YALE BIS634 Final Project

The project is mainly about analyzing the COVID-19 vaccine tweets dataset.

This local Web application is built with Python flask, along with some HTML and CSS.

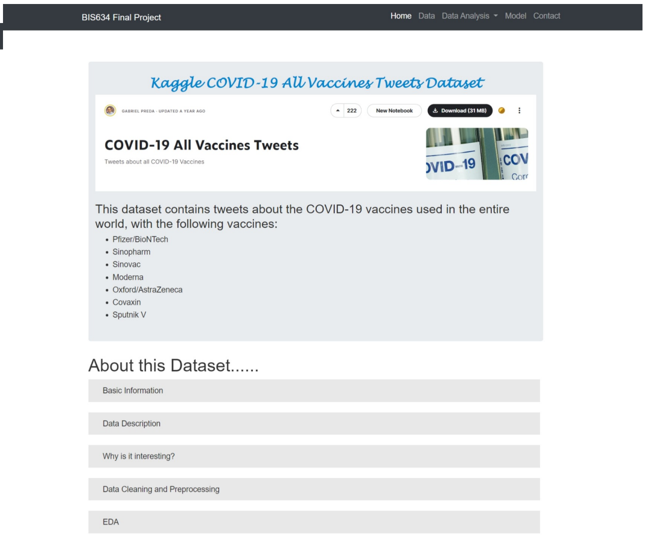
There are 7 app routes

- @app.route('/'): Home Page
- @app.route('/data'): Data Page.
- @app.route('/wordcloud'): Word Cloud Analysis Page.
- @app.route('/timeseries'): Timeseries Analysis Page.
- @app.route('/model_page'): Model Page without Output Result.
- @app.route('/model_result'): Model Page with Output Result.
- @app.route('/contact'): Contact Page

and 7 APIs.

- @app.route('/wordcloud_api/<string:vax_name>'): Returns JSON-encoded data generating a word cloud given a vaccine brand.
- @app.route('/wordcloud_api/senti/<string:senti>'): Returns JSON-encoded data generating a word cloud given a sentiment.
- @app.route('/wordcloud_api/country/<string:country>'): Returns JSON-encoded data generating a word cloud given a country name.
- @app.route('/timeseries_api/varianceVStime/<string:vax_name>'): Returns JSON-encoded data generating a timeseries graph given a vaccine brand.
- @app.route('/timeseries_api/varianceVStime_Country/<string:country>'): Returns JSON-encoded data generating a timeseries graph given a country name.
- @app.route('/timeseries_api/SDT/<string:senti>'): Returns JSON-encoded data generating a bar graph given a sentiment.
- @app.route('/model_api/<string:user_text>'): Returns JSON-encoded data generating the predicted results from the model given a user input.

Figure 10: Home Page



Kaggle COVID-19 All Vaccines Tweets Dataset

COVID-19 All Vaccines Tweets

Tweets about all COVID-19 Vaccines

This dataset contains tweets about the COVID-19 vaccines used in the entire world, with the following vaccines:

- Pfizer/BioNTech
- Sinopharm
- Sinovac
- Moderna
- Oxford/AstraZeneca
- Covaxin
- Sputnik V

About this Dataset......

Basic Information

Data Description

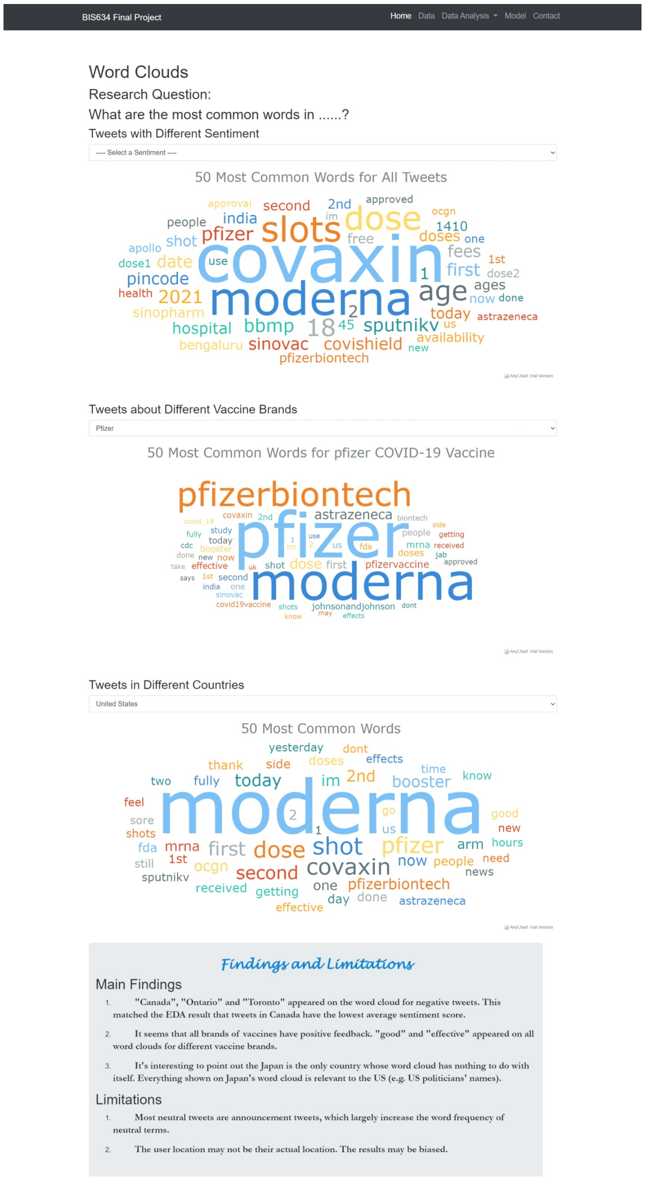
Why is it interesting?

Data Cleaning and Preprocessing

EDA

Figure 11: Data Page

- @app.route('/wordcloud'): Word Cloud Analysis Page. Displays the interactive word clouds and the findings.
- @app.route('/timeseries'): Timeseries Analysis Page. Displays the interactive timeseries plots and the findings.



Word Clouds

Research Question:

What are the most common words in ......?

Tweets with Different Sentiment

Tweets about Different Vaccine Brands

Tweets in Different Countries

Findings and Limitations

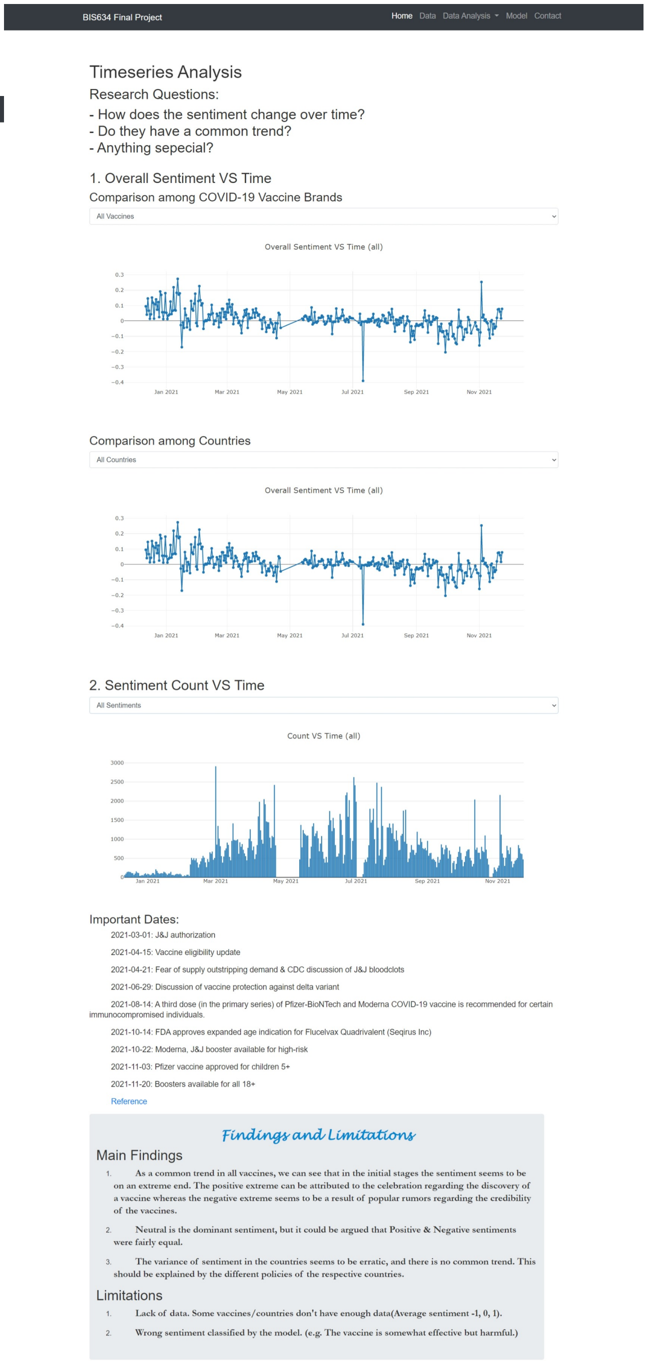
Main Findings

1. "Canada", "Ontario" and "Toronto" appeared on the word cloud for negative tweets. This matched the EDA result that tweets in Canada have the lowest average sentiment score.
2. It seems that all brands of vaccines have positive feedback. "good" and "effective" appeared on all word clouds for different vaccine brands.
3. It's interesting to point out the Japan is the only country whose word cloud has nothing to do with itself. Everything shown on Japan's word cloud is relevant to the US (e.g. US politicians' names).

Limitations

1. Most neutral tweets are announcement tweets, which largely increase the word frequency of neutral terms.
2. The user location may not be their actual location. The results may be biased.

Figure 12: Word Cloud Analysis Page



Timeseries Analysis

Research Questions:

- How does the sentiment change over time?
- Do they have a common trend?
- Anything sepecial?

1. Overall Sentiment VS Time

Comparison among COVID-19 Vaccine Brands

Comparison among Countries

2. Sentiment Count VS Time

Important Dates:

2021-03-01: J&J authorization

2021-04-15: Vaccine eligibility update

2021-04-21: Fear of supply outstripping demand & CDC discussion of J&J bloodclots

2021-06-29: Discussion of vaccine protection against delta variant

2021-08-14: A third dose (in the primary series) of Pfizer-BioNTech and Moderna COVID-19 vaccine is recommended for certain immunocompromised individuals.

2021-10-14: FDA approves expanded age indication for Flucelvax Quadrivalent (Seqirus Inc)

2021-10-22: Moderna, J&J booster available for high-risk

2021-11-03: Pfizer vaccine approved for children 5+

2021-11-20: Boosters available for all 18+

Reference

Findings and Limitations

Main Findings

1. As a common trend in all vaccines, we can see that in the initial stages the sentiment seems to be on an extreme end. The positive extreme can be attributed to the celebration regarding the discovery of a vaccine whereas the negative extreme seems to be a result of popular rumors regarding the credibility of the vaccines.
2. Neutral is the dominant sentiment, but it could be argued that Positive & Negative sentiments were fairly equal.
3. The variance of sentiment in the countries seems to be erratic, and there is no common trend. This should be explained by the different policies of the respective countries.

Limitations

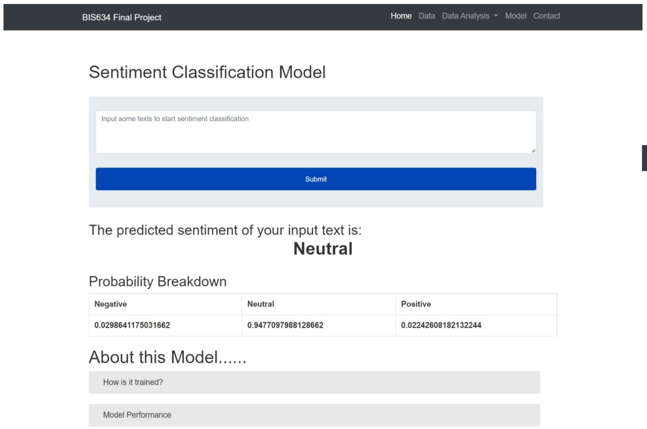
1. Lack of data. Some vaccines/countries don't have enough data(Average sentiment -1, 0, 1).
2. Wrong sentiment classified by the model. (e.g. The vaccine is somewhat effective but harmful.)

Figure 13: Timeseries Analysis Page

- @app.route('/model_page'): Model Page without Output Result. Displays the input box and the model information.
- @app.route('/model_result'): Model Page with Output Result. Displays the input box, model output,

and the model information.

- @app.route('/contact'): Contact Page. Displays my contact information and Web application source code.



Sentiment Classification Model

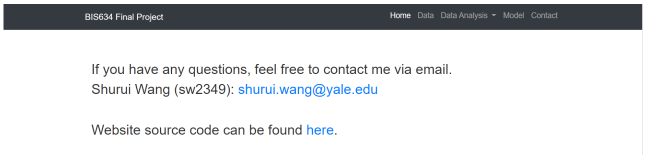
The predicted sentiment of your input text is:
Neutral

Probability Breakdown

Negative	Neutral	Positive
0.0298641175031662	0.9477097988128662	0.02242608182132244

About this Model......

Figure 14: Model Page/Result Page



If you have any questions, feel free to contact me via email.
Shurui Wang (sw2349): shurui.wang@yale.edu

Website source code can be found here.

Figure 15: Contact Page

6 Limitations

1. Due to the lack of data, some countries/vaccines have a relatively low sample size which leads to an unusually high/low average sentiment score. Thus, the results may be misleading and inconclusive.

2. Since the sentiment labels are predicted using the model, and the model accuracy is about 80%, it is possible that the labels are wrongly marked for some tweets. For example, some neutral tweets are not necessarily neutral. This may lead to analysis bias.

3. Since Twitter is not the only social media platform for people to leave feedback, the results cannot represent the overall reputation of a specific vaccine in a specific country. Also, since the data is outdated, the analysis results only represent 2021, not now.

7 Unexpected Difficulties

1. Since I'm not familiar with JavaScript and JQuery, it is extremely hard for me to make the interactive plots work on my Website. I have no idea how to get the user input/selected values from dropdown menus and implement them on my website. Also, the tutorials online match my needs are limited.

2. Another difficulty is the unexpectedly long time to train the machine learning model and predict the results. A minor mistake in data cleaning will lead to re-running all the code which was a painful and time-consuming experience.

Finally, I overcome all difficulties and made everything work.

8 Appendix

Here are links to my data sources:

1. COVID-19 All Vaccines Tweets Dataset.

https://www.kaggle.com/datasets/gpreda/all-covid19-vaccines-tweets

2. Complete Tweet Sentiment Extraction Data.

https://www.kaggle.com/datasets/maxjon/complete-tweet-sentiment-extraction-data