Vidur Sinha, Rahul Banerjee
Project Code: https://github.com/vsinha1124/LIN373N_Final_Project

LIN 373N - Final Report

INTRODUCTION

As more COVID-19 vaccines begin to roll out, many people want to know which one to take and the possible risks associated with each vaccine. Many people have experienced side effects from the vaccines such as fever, headaches, muscle pain, and tiredness, particularly after the second dose. As a result, there have been thousands of tweets containing opinions, research, and personal experiences that people have with COVID-19 vaccines. Gathering insights from these tweets, particularly sentiment analysis, can help people with their decision on which vaccine to take. Exploring this data can also help people better prepare for the side effects of the vaccines. For instance, if there are many tweets which report fever the day after the second dose, people can prepare to take a day off of work. We also plan to explore the timeline of various news articles/posts regarding the vaccines to see their effects on the sentiment of the tweets. This will help provide us with another layer of understanding of the tweet-data and give us insights into how people react on twitter to various types and sources of news.

DATA

Since we will be using an unlabeled COVID-19 vaccine tweets dataset, we need some way to train a model to predict their sentiments. For this, we will be using the Sentiment140 dataset to train our models. This is a labeled dataset containing over 1,600,000 tweets.

The primary dataset we will be using for the COVID-19 tweets is the "All COVID-19 Vaccines Tweets" from Kaggle. It is updated daily and contains both the tweets and metadata about them. We plan to analyze the sentiment toward the different available vaccines (Pfizer, AstraZeneca, etc.) in different parts of the world. We also plan to use data on the COVID-19 world vaccination progress to find a relationship between how many people have been vaccinated and people's attitudes toward vaccines.

METHODOLOGY

Before we trained our models, we had to preprocess the Sentiment140 tweets. This included removing things such as emojis, urls, punctuation, symbols, and stopwords. We also decided to go with the TfidfVectorizer as our method of feature extraction, and we used both unigrams and bigrams for this. We also tried out BERT as a replacement for TfidfVectorizer, but we decided against using it for our final models. This is because training BERT took far too long— we were not seeing a significant enough boost in accuracy to justify waiting days for it to finish.

The models we attempted were logistic regression, random forest, naive Bayes, linear SVC, AdaBoost, and LSTM neural networks. We compared these models on a 100,000 sample subset of Sentiment140 since the original 1.6 million tweets took far too long to train on. Our baseline was to use the models with the default hyperparameters, and we used sklearn's GridSearchCV to tune them. After comparing the models using 10-fold cross-validation, we found that logistic regression and the LSTM performed the best with an accuracy of around 76% on the subset (10 fold cross validation). However, the logistic regression model took far less time to train on, so we figured training it on the entire Sentiment140 dataset before testing on the COVID-19 tweets dataset would be the most practical choice. For the COVID-19 tweets dataset, we broke down our analysis based on the location of the tweet, the date it was tweeted, and some keywords in the tweets themselves.

RESULTS

After our testing, we determined that Logistic Regression with 'newton-cg' solver was the best combination to use— with a 79% accuracy when tested on the entire Sentiment140 dataset. We used this model to classify the tweet sentiments for the COVID-19 vaccine tweet dataset.

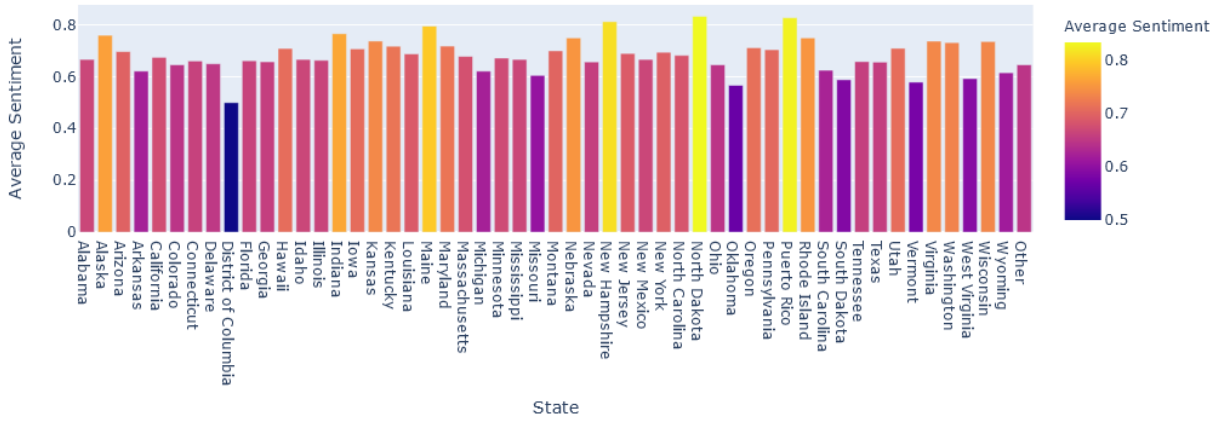

Figure 1: USA States Sentiments

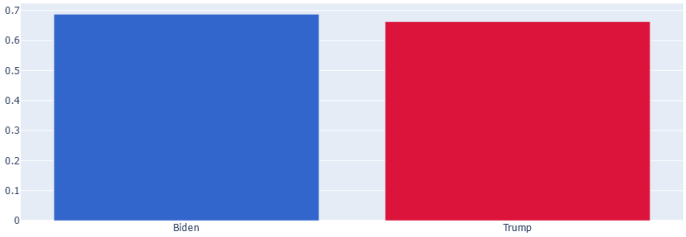

Figure 2: Blue vs Red States Sentiments

In our analysis, our group decided to study how different states in the US perceive COVID-19 vaccines in general. Figure 1 shows our results, and we see that states like North Dakota and Maine appeared to have the highest overall sentiments towards the vaccines. We were not expecting these results, but we think that this phenomenon is because these states are relatively well vaccinated compared to the national average. We also see that Washington DC has very low sentiment, but we believe that this figure is somewhat skewed by the fact that many reporters are based out of Washington DC, so negative news pertaining to vaccines and vaccine distribution is more likely to come from there. We also divided the states into who they voted for in the 2020 Presidential Election, as seen in Figure 2. We observed an approximately 0.02 higher sentiment in states that voted for Biden compared to states that voted for Trump. This makes sense, as conservative news outlets have covered the COVID-19 pandemic very differently than more liberal ones.

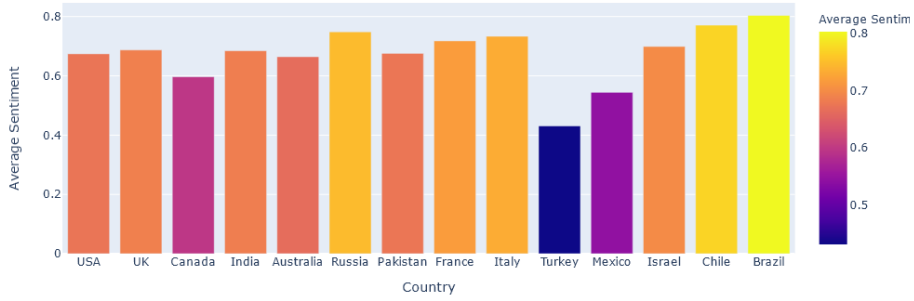

Figure 3: Sentiments by Country

Our team also studied tweets from 14 countries around the world. We picked these countries based on how badly they have been hit by COVID-19, and how well they have handled their vaccination efforts. We notice that Israel and Chile have among the highest sentiments, which makes sense given that they have some of the highest vaccination rates in the world. We also notice some unexpected results in the data. Brazil and Russia have very high vaccine sentiments, but most news accounts describe very dire situations in those countries. One reason for this discrepancy is selection bias. These are English tweets, and the primary language in those countries is not English,

so these tweets are probably not representative of the true situations there. Moreover, the way these tweets were collected is through a user's use of certain key COVID-19 vaccine-related hashtags. So if a user did not include a hashtag, then their tweet was not included in this dataset. This is a potential source of error in our analysis.

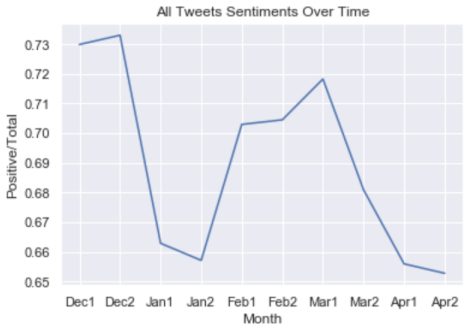

Figure 4: Overall Sentiments over time

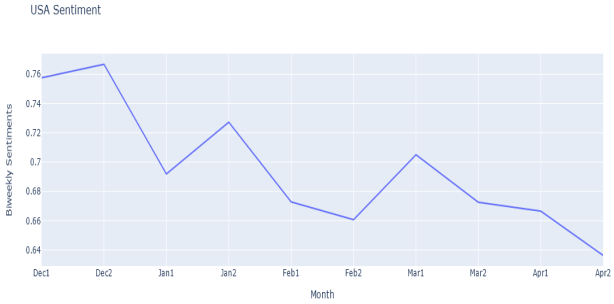

Figure 5: USA Sentiments over time

In Figure 4, we can see how the average sentiment of the entire dataset has changed over time. Sentiments began high as expected, with lots of hope of life returning to normal, with a strong correction due to restrictions and issues with vaccine rollout early this year. Then sentiment began to rise again in the spring with many people receiving their vaccines, and lastly there was again a sharp decline in April. We assume this is due to a low sentiment in individuals who have not gotten the vaccine due to their personal views about the COVID vaccines or vaccines in general. In Figure 5, we see that there has been a slow decline in the sentiments toward the vaccines in the US.

US Vaccines Analysis

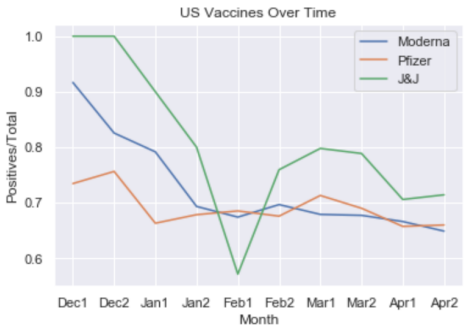

Figure 6: USA Vaccine sentiments over time

In Figure 6, we can see how the sentiments toward each of the three US vaccines have changed over time. As expected, these vaccines began with their highest sentiments in early december and have slowly declined since then. Additionally, there were reports of patients getting blood clots from the Johnson & Johnson vaccine which explains the sharp decline in the Johnson & Johnson vaccine sentiment in February.

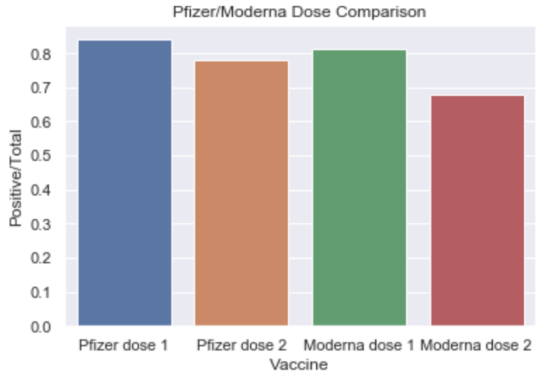

Figure 7: Moderna vs Pfizer dose comparison

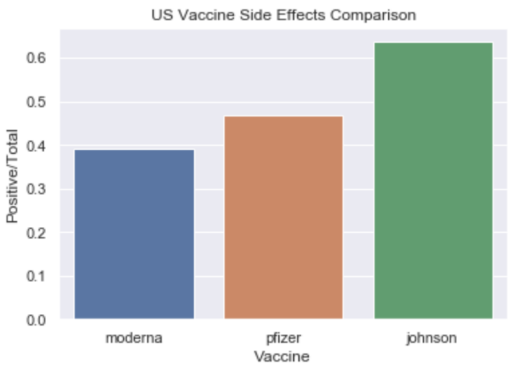

Figure 8: US vaccine side effects comparison

In Figure 7, we have analyzed the differences in sentiment between the two doses of the Pfizer and Moderna vaccines. This was done by taking the intersections of the indexes which contain the vaccine name and either “first dose” or “second dose.” We then took the average sentiments of the tweets at these indexes. The differences in sentiment between the two vaccines and each of their doses reflects the general sentiments we have heard from people we know who have taken the vaccines. Generally, the side effects are worse for the second doses of the vaccines, but Moderna usually has worse side effects.

In Figure 8, we can see the differences in sentiment for all three vaccines overall in regard to their side effects. Again, as expected, Moderna has worse side effects than Pfizer, and Johnson & Johnson has the highest sentiments in regard to side effects, indicating that it induces the least amount of side effects.

Indian and European Vaccines Over Time

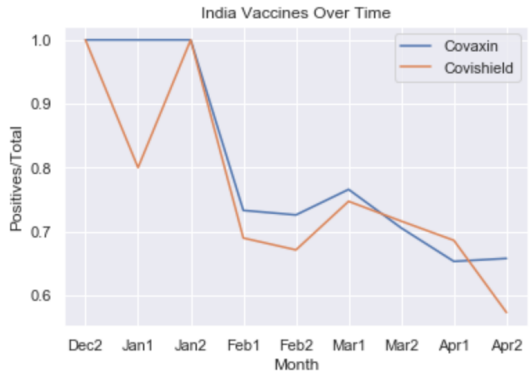

Figure 9: India’s vaccine sentiments

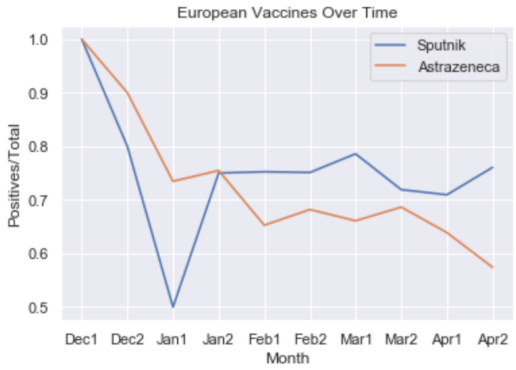

Figure 10: European vaccine sentiments

With the recent increase in COVID-19 cases in India, we decided to see how the sentiment toward their two vaccines have changed over the past few months. As expected, there has been a drastic decline in the sentiment towards their vaccines, as seen in Figure 9. Additionally, we noticed the higher sentiment toward the Covaxin vaccine, which, unlike Covishield, was developed entirely by India. Covishield has a slightly higher efficacy over Covaxin, so we suspect the higher sentiment is due to national pride toward their vaccine.

We also analyzed how the sentiments toward two European vaccines, Sputnik V and Astrazeneca have changed over time. With higher efficacy rates, Sputnik V has generally had higher sentiments than Astrazeneca, as seen in Figure 10. There were suspicions in January that Russia’s reports on the effectiveness of the Sputnik V vaccine were faked or exaggerated, which explains the sharp decline in sentiment during that time.

TEAM STRUCTURE

For training on Sentiment140, Rahul tested BERT, Logistic Regression, Adaboost, Multinomial naive Bayes, Gaussian Process Classifier. Vidur tested Neural Networks (LSTM), Random Forest, Logistic Regression. For COVID vaccine tweet analysis, Rahul performed sentiment analysis by country/location (Twitter user location). Vidur performed sentiment analysis by vaccines (Tweet content) and time (time/date Tweet was posted).

REFERENCES

Go, A., Bhayani, R. and Huang, L., 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford, 1(2009), p.12*.

Kruspe, Anna et al. 2020. Cross-language sentiment analysis of European Twitter messages during the COVID-19 pandemic. *Association for Computational Linguistics.*

Preda, Gabriel. All COVID-19 Vaccines Tweets. 2020. https://www.kaggle.com/gpreda/all-covid19-vaccines-tweets. Online; accessed March 16, 2021.