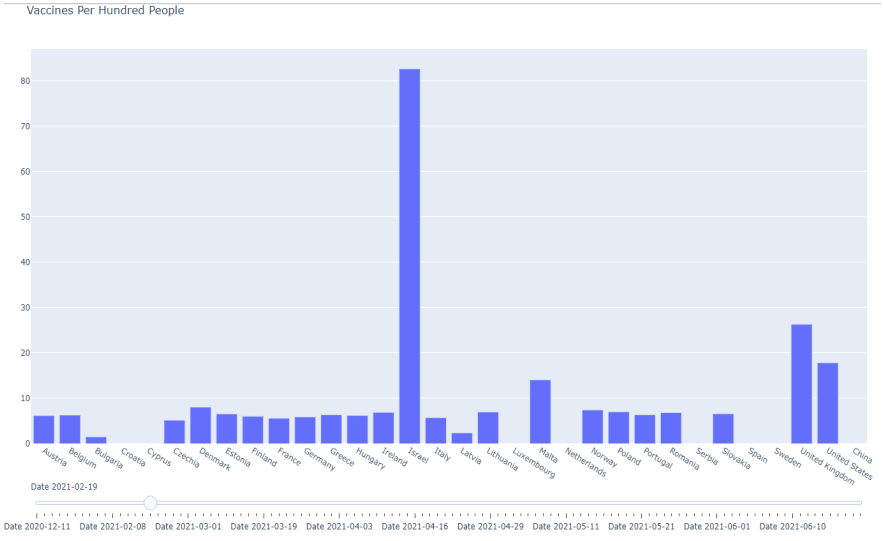
1. **Which tasks have been completed?**

We first loaded the datasets and read the CSV files. The three datasets include tweets about “all brands of vaccines dataset,” “total vaccination for COVID-19 in the world dataset,” and “the extraction of tweets sentiment dataset.” Then, we preprocessed the datasets. For our language model, the only input we need is the ‘tweet text.’ Fastai can do the text preprocessing for us, and at the same time, we removed things from the ‘text column,’ such as URLs, emojis, and rows that are missing sentiment. Since there are two datasets, “all brands of vaccines” and “the extraction of tweets sentiment,” we cleaned the text and combined the two datasets with only two columns left: ‘text column’ and ‘sentiment column.’ The ‘sentiment column’ tells us about whether the corresponding text is positive, negative, or neutral.

In addition, we transformed negative, positive, and neutral sentiments to 0, 1, and 2 correspondingly for further modeling. Next, we tokenized the texts and performed lowercase, punctuation removal, small token removal, stop words removal, lemmatization, and stemming to them. After all these steps, we got a dataset with four columns: original text, sentiment, final text, and text tokens. Afterward, in order to get a general sense of the sentiment prediction task, we firstly applied Naive Bayes to the dataset and obtained an accuracy of 0.6260. Then, we applied another baseline model to the dataset, XGBoost, which gave us a little higher accuracy of 0.6964 after rounds of parameters tuning.

Furthermore, we added visualization to the vaccination progress data where we illustrated the correlations among multiple factors such as vaccination rate, total vaccinations, and daily vaccinations among many different countries.



In the above sample figure, we made an interactive plot to show the vaccine doses per hundred people among different countries. Here we made a date as a slider bar to dynamically display the movement of the statistics. We also made interactive plots to show the movement of vaccination rate, daily vaccinations, and fully vaccinated people in given time duration. With the help of these plots, we can better illustrate the vaccination progress data and better analyze the correlation between vaccination progress and Twitter vaccine sentiments.

2. **Which tasks are pending?**

Firstly, as mentioned before, the accuracies of sentiment prediction from baseline models were not high enough. The highest accuracy was only 0.6964. Therefore, we plan to self-learn some deep-learning-based models and then find the best one for our application to achieve higher accuracy for our prediction task.

Secondly, we are still analyzing the timelines added into the datasets for each type of COVID-19 vaccine and then trying to visualize the relationship between vaccination and countries. For example, what type of vaccination is used in different countries, which vaccine is the most widely used, and the relationship between the number of vaccinations daily/per people/per country and try to visualize the progress of vaccinations evolved from each country. We will further improve the visualizations of vaccination progress data, implement visualizations of the predicted Twitter sentiment data, analyze these results, and find some correlations between Twitter sentiment and real vaccination progress.

3. **Are you facing any challenges?**

Right now, the challenge we are facing includes interpreting sarcasm on some occasions. Expressing negative sentiment using backhanded compliments makes it hard for sentiment analysis tools to detect the true context of what the response is actually implying. Besides, social media contents, just like the tweets on Twitter, are text-based, which means that they are inundated with emojis. While NLP tasks can extract text from even images, emojis are a language in itself. Most of the time, we treat emojis as special characters that are removed from the data during data mining, but this may cause the loss of data insights and accuracy.