# Team: Chirps

**Team Members:** Ashika Anand Babu, Subarna Chowdhury Soma, Vishnu Vardhan Reddy Yeruva

## Project Name: Public opinion and sentiment analysis on the COVID-19 vaccine using Twitter data

## Data Science Approaches and Algorithms:

1. **Sentiment Computations:** For sentiment analysis of Tweets, we have used two libraries.

    **1.1** At first we have used Vader library to compute sentiment and sentiment score

    **1.2** Then we have used TextBlob library to compute Polarity and Subjectivity from COVID-19 vaccine twitter dataset

2. **Sentiment Classification/Prediction Model:** To build the classification model, we have used several approaches and algorithms.

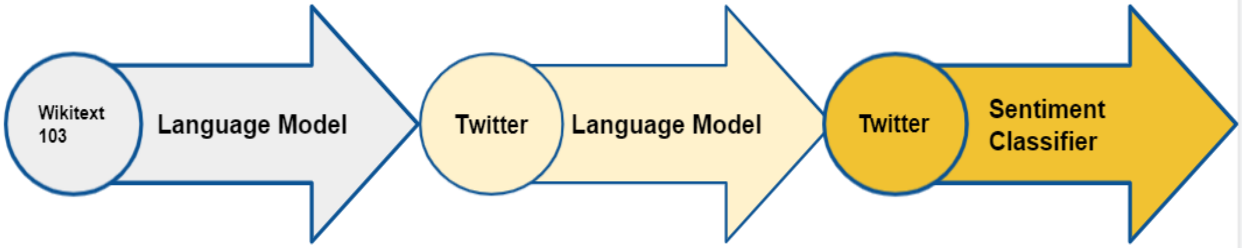
    **2.1 Multiclass Classification Algorithms:** First of all, we have used several machine learning algorithms to build classification/prediction models. These models are multiclass classification models as there are three sentiments to classify: Positive, Negative and Neutral. Following are the used algorithms with their accuracy.

| No. | Algorithms | Accuracy |
|---|---|---|
| 1 | Logistic Regression | 0.893236 |
| 2 | Stochastic Gradient Decent | 0.881183 |
| 3 | Random Forest | 0.872281 |
| 4 | CatBoost | 0.869907 |
| 5 | Naive Bayes | 0.769763 |
| 6 | XGBoost | 0.724382 |

**2.2 BinaryClass Classification Algorithms: T**o improve the accuracy, we have converted our multiclass data into binary class data. For that 'Neutral' class is also mapped to the 'Positive' class. As expected, for binary class classification improved a lot for each algorithms

| No. | Algorithms | Accuracy |
|---|---|---|
| 1 | Logistic Regression | 0.930080 |
| 2 | Stochastic Gradient Decent | 0.923756 |
| 3 | Random Forest | 0.923346 |
| 4 | CatBoost | 0.922912 |
| 5 | Naive Bayes | 0.881503 |
| 6 | XGBoost | 0.863195 |

**2.2 Deep Learning Approach (Transfer learning in fastai - the ULMFiT):** We have two kaggle datasets of tweets. Therefore we have built a language model using a transfer learning technique for NLP called 'Universal Language Model Fine-Tuning' (ULMFiT).. Our intention is to explore the advanced side of data science. We have used the fastAI library here. Following are the steps taken:

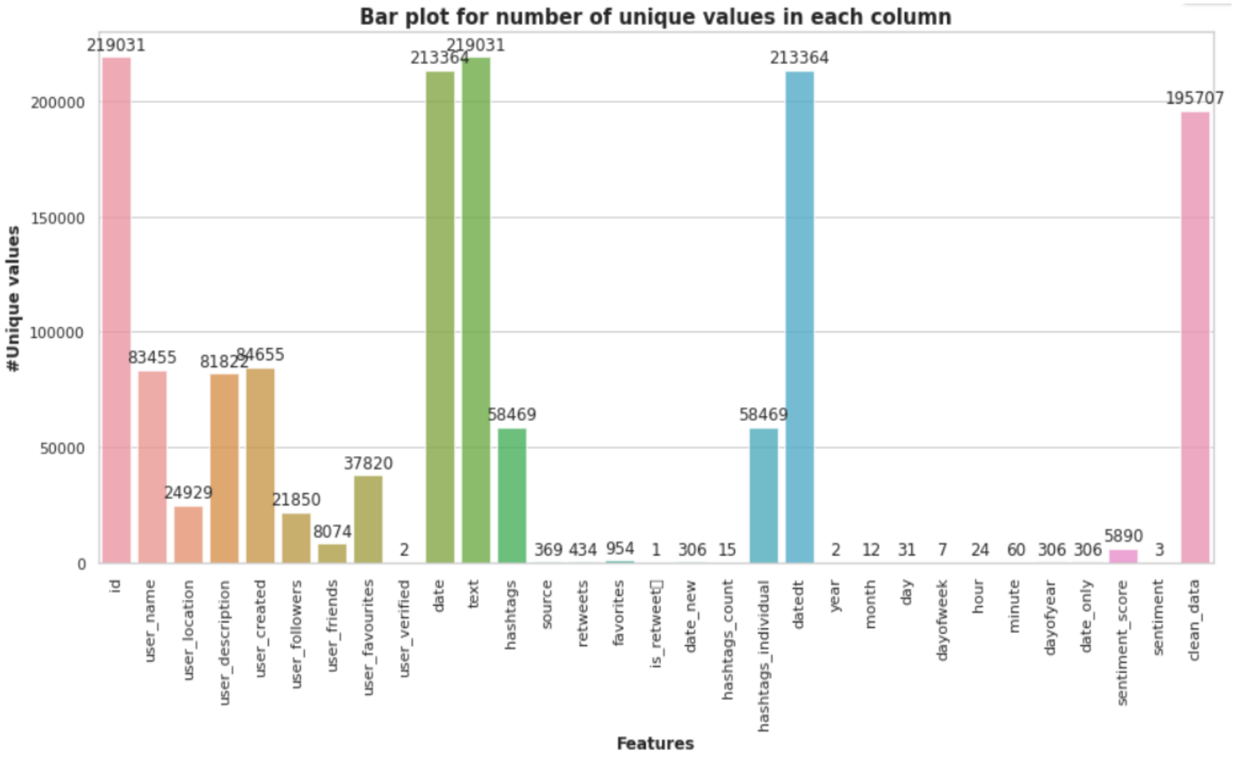


**a.** We already have a trained language model to predict the next word in a sentence.

**b.** Then we have fine-tuned the language model for our sentiment classification purpose. Wikipedia English can be understood by pre-trained language models, but Twitter English is a little different. By taking the information from the Wikipedia model and applying it to our Twitter datasets, we can find a language model that correctly predicts the next word in a tweet. For this purpose, we have merged both of our twitter datasets.

c. Finally, our fine-tuned classification model can identify tweet's sentiment using the pre-trained language model. The idea is to build a classifier that understands both positive and negative sentiment based on what our language model already knows about Twitter English. It would be very time consuming and difficult for a classifier to be trained without a pre-trained model. Therore, we have used this advanced deep learning approach to classify and analyse sentiment in the COVID-19 vaccine tweets and compare changes in sentiment for different vaccine types. It takes a long time to train the model, so we have trained it in very small iterations. With that the accuracy is 75%. The accuracy can be improved by training it for further epochs with more finetuning.

## Features Used



The main features we have explored and worked on are:

1. User location
2. Hashtag
3. Date
4. And the TWEET itself

Features Derived:

1. Vaccine type
2. Count of tweets
3. Polarity
4. Sentiment score

### User location

From our labelled dataset, we obtain the user location and perform location based vaccine type derivation from the data

### Hashtag

From the hashtag in the tweet, we separate the hashtag from the tweet, find all the various spellings and various representations to identify the vaccine type. We also count the number of tweets based on the type of vaccine

### Date

For analysis over time, we use the date, month, year feature in the dataset

### Tweet text

The tweeted text is used for performing sentiment analysis by identifying the polarity of the tweet, sentiment score of the tweet.

### Vaccine type

Vaccine type is a derived feature obtained from the hashtags using one hot encoding. Using this feature we are able to separate the tweet based on the type of vaccine

### Count of Tweets

We count the tweets based on vaccine, location and polarity for representing how the reaction of people have varied over time and type of vaccine taken

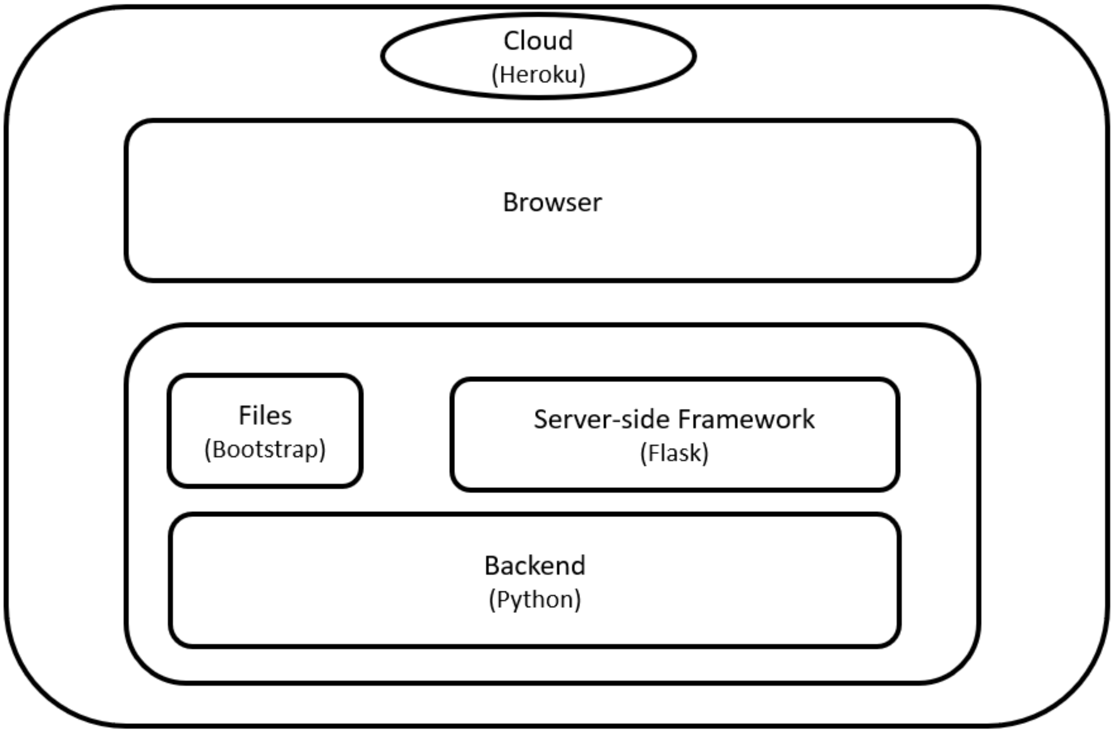
### Polarity

We obtain the polarity of the tweet for identifying if the tweet was neutral, positive or negative.

### Sentiment Score

We derive the sentiment score from the tweet itself to analyse the inclination of the tweet.

## Client Side Design



### Cloud

Application deployed into the cloud is the shell of entire client centric development which makes it as a seamless interaction between user and application

### Browser

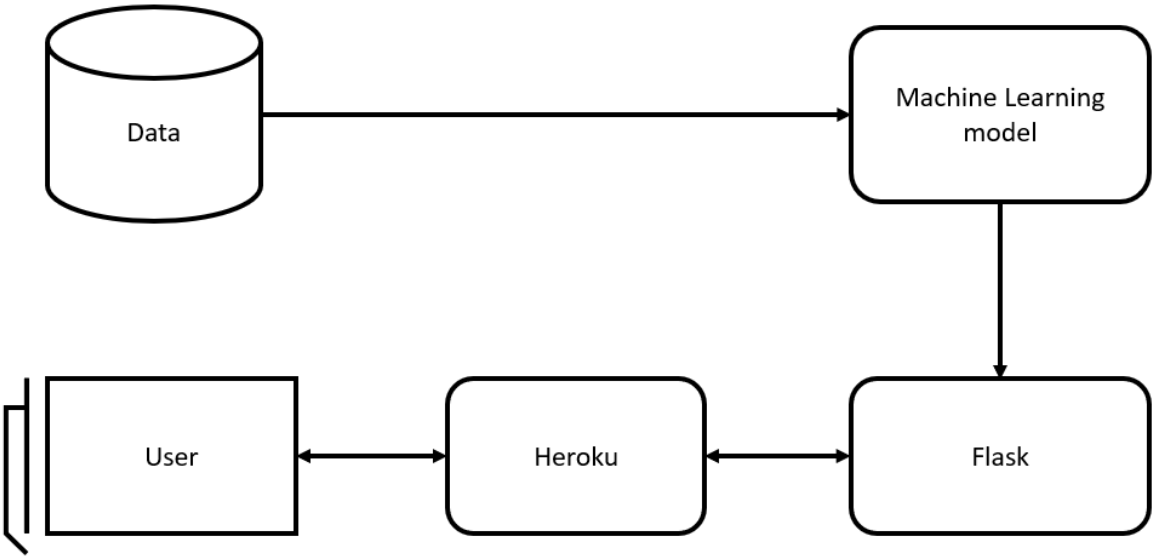
Most common and feasible component for a web application and Edge, Opera, Firefox, Chrome, Safari, Safari mac, Firefox mac, Chrome mac are the major browsers while writing this document. As you can see, trying to build and test everything is difficult. Each browser has its own subtle nuances different in browser security, default font sizes, borders etc. All these issues can be overcome with the changes required in the programming.

### Programming

Programming with required components whether it can be frontend or backend captivates the users experience.
To make it a more user friendly and interactive interface it always moves in a frontend’s direction. Frontend is a key that always captures customer’s satisfaction to yield more income.

Backend is always a hidden gem that performs it’s actions to strengthen the application and makes it more efficient.

## Model Deployment



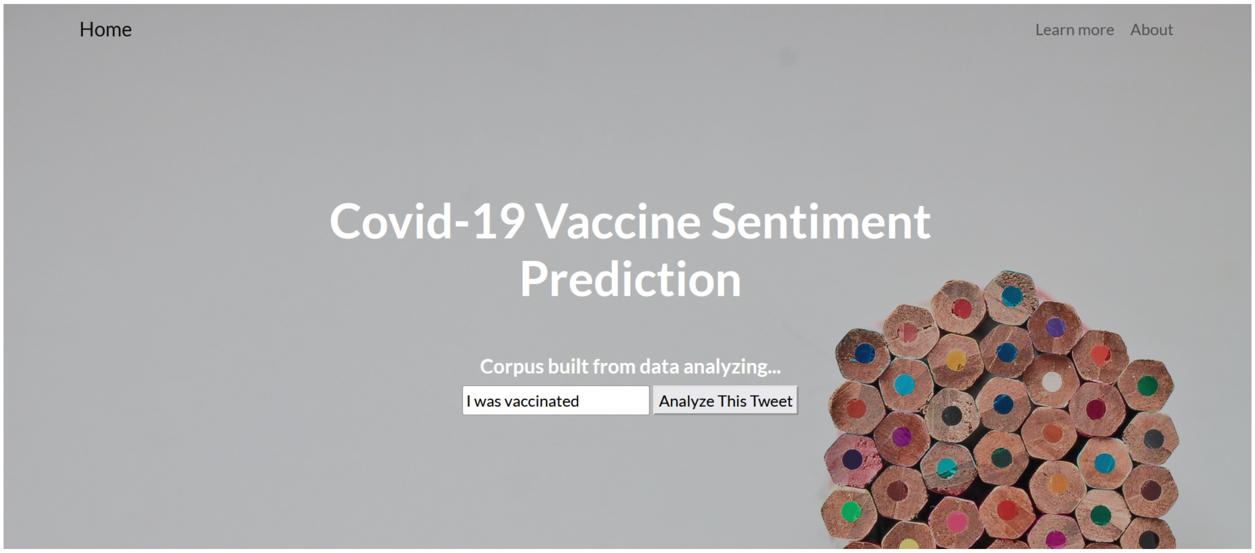
Fetching Training data ----------------> training model --------------------> Evaluating model -------------------> Model Endpoint

Fetching and training data from the trusted sources and then building a model to attain a classifier and using that pickle file/ providing an interface by making it a supervised approach takes it further by gathering inputs from users and to make a predictions with an accuracy of 88.5% as high as possible for a SGD multiclass classifier in this spectrum looks an efficient build. Deploying the same into the cloud by compressing slug size without any compromises is an added advantage. We have used this model to make COVID-19 vaccine related tweet sentiment prediction.

## Client Side Application: Web App

## Link: https://covid-vaccine-sentiment-pred.herokuapp.com/

To make Interface more interactive, scalable and enjoyable we have picked Flask and Bootstrap as our designers. As Flask is a micro-framework i.e with little to no dependencies to external libraries is the reason we picked it over Django and it is light, there are little dependency to update and watch for security bugs

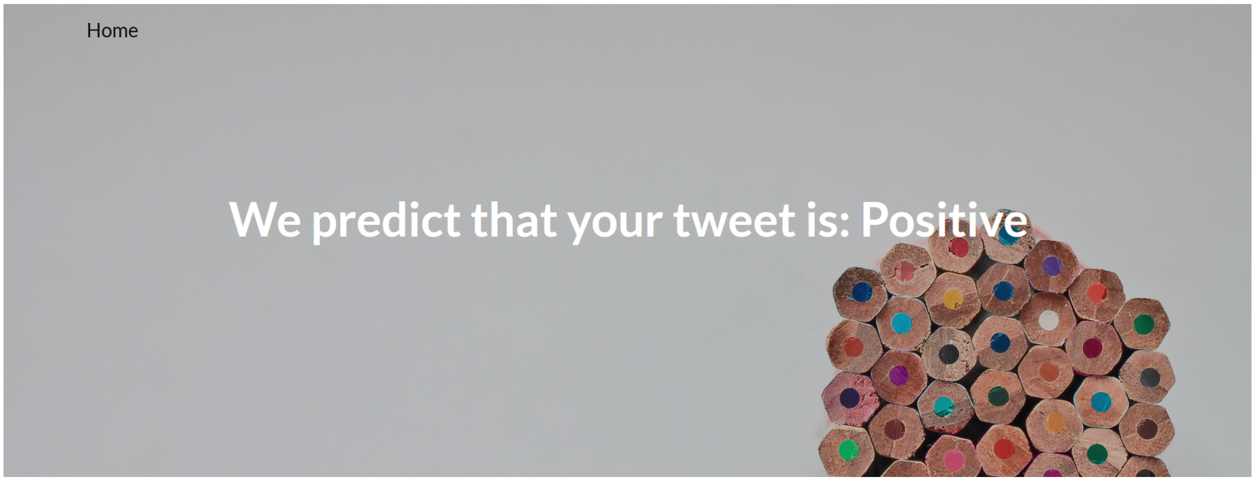


Covid-19 Vaccine Sentiment Prediction

Corpus built from data analyzing...

I was vaccinated

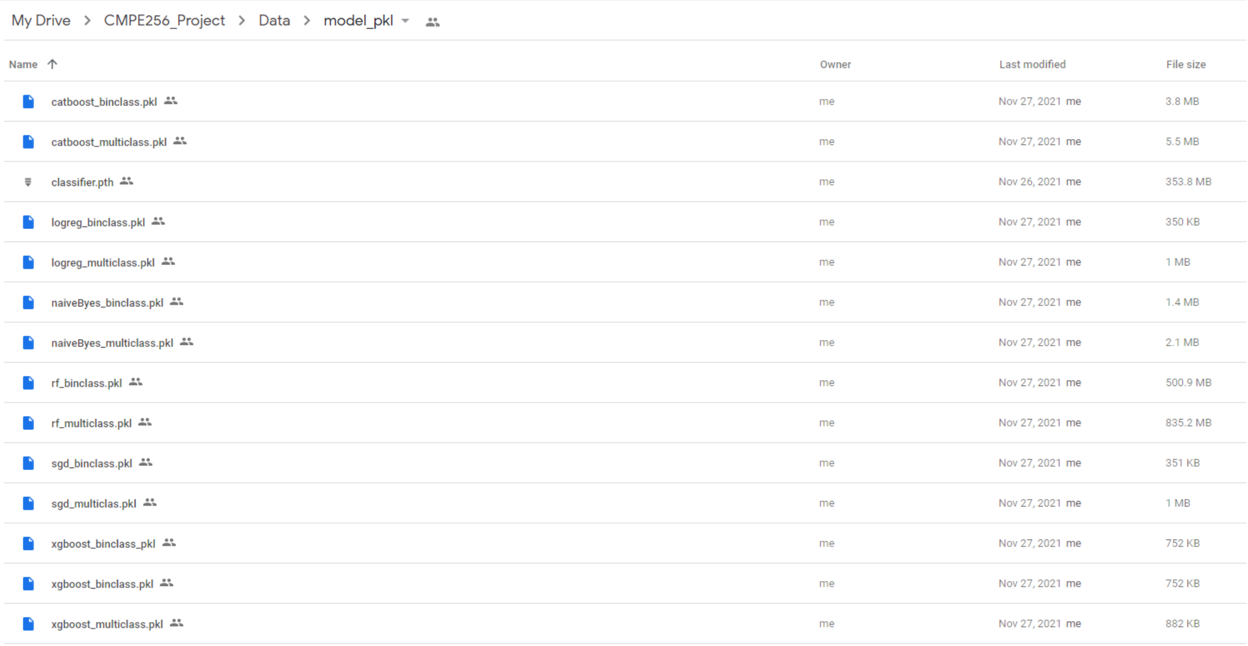
We have crafted it to be simple yet elegant to use. The straight forward interface which has sentiment analyzer on the home screen and redirects to the result page with a simple click.



We predict that your tweet is: Positive

## PKL Files:

We have 14 PKL files generated from multiclass, binary class classification and language model. In the canvas we are submitting the best one SGD Model that we have used for inference. Rest of the PKL files can be found here: https://drive.google.com/drive/folders/1ZpqmD4J5b4ZmEYAGf-e2VkNY3uSvcko4?usp=sharing



My Drive > CMPE256_Project > Data > model_pkl

| Name ↑ | Owner | Last modified | File size |
|---|---|---|---|
| catboost_binclass.pkl | me | Nov 27, 2021 me | 3.8 MB |
| catboost_multiclass.pkl | me | Nov 27, 2021 me | 5.5 MB |
| classifier.pth | me | Nov 26, 2021 me | 353.8 MB |
| logreg_binclass.pkl | me | Nov 27, 2021 me | 350 KB |
| logreg_multiclass.pkl | me | Nov 27, 2021 me | 1 MB |
| naiveByes_binclass.pkl | me | Nov 27, 2021 me | 1.4 MB |
| naiveByes_multiclass.pkl | me | Nov 27, 2021 me | 2.1 MB |
| rf_binclass.pkl | me | Nov 27, 2021 me | 500.9 MB |
| rf_multiclass.pkl | me | Nov 27, 2021 me | 835.2 MB |
| sgd_binclass.pkl | me | Nov 27, 2021 me | 351 KB |
| sgd_multiclas.pkl | me | Nov 27, 2021 me | 1 MB |
| xgboost_binclass_pkl | me | Nov 27, 2021 me | 752 KB |
| xgboost_binclass.pkl | me | Nov 27, 2021 me | 752 KB |
| xgboost_multiclass.pkl | me | Nov 27, 2021 me | 882 KB |

## Project Colabs:

We have three collabs to do the pre processing, exploratory analysis, finding features, sentiment computation and finally model generation.

1. 1.CMPE256_Covid19_VacTweet_PreProcessing_Sentiment_Analysis_(CHIRPS).ipynb
2. 2.CMPE256_Covid19_VacTweet_ExploratoryAnalysis_(CHIRPS).ipynb
3. 3.CMPE256_Covid19_VacTweet_sentiment_classification_model_(CHIRPS).ipynb