Information Retrieval – CSE 535 – Final Project

Team – Status_200

Introduction

The aim of the final project was to create an end-to-end web search engine along with an analytics dashboard. The search engine used covid and vaccine twitter tweets for indexing and used the Okapi BM 25 model to retrieve relevant tweets for a given query. The backend was built in flask, as for the analytics dashboard, we used Plotly.js for innovative story telling and interactive data plots.

Methodology:

The overall structure of the search is as follows:

Technologies/frameworks used in frontend: Plotly.js, CSS and HTML.

Technologies/frameworks used in backend: Flask, Solr, NLTK, Keras, Flair

The User Experience Flow:

Upon entering the website, the user is presented with a clean interface to enter a query. This is rendered via HTML and CSS. Upon entering the query, the query is sent to the flask server, where the following happens:

1. The Flask Engine performs stop word removal for all languages.
2. The cleaned query is then sent to Solr to retrieve the appropriate results.
3. Once the "all" the results are retrieved, we compute the global statistics and create a view for the search query result page. The following graphs are generated based on the query and filters:
    a. Tweet Count of POIs and General Population for the query.
    b. Language tweet count.
    c. Country-wise tweet count.
    d. Tweet count for POIs whose tweets have been retrieved.
4. We also present the following filters to the user to narrow down the tweets they are looking for:
    a. Language wise tweets.
    b. Country wise tweets.
    c. POI name wise tweets.
5. Once the graphs are generated, we then take the top 40 tweets to present to the user.
    a. If POI tweets are available, the top tweets are dominated by those, followed by the general population tweets. If we have 40 tweets by POIs, all tweets will be from POIs.
6. The following components are displayed as part of the tweet:
    a. User-name: If available, otherwise Twitter User (POI names are always available, hence POI name will always be present).
    b. If the tweet is by a POI, then a hyperlink to the tweet will be present.
    c. Whether a user is verified or not.
    d. The tweet text.
    e. The tweet sentiment and the probability of the sentiment.
    f. Best positive reply if present, best negative reply if present.

This summarizes the search page view.

Moving on the Analytics dashboard (Overview Page), we have precomputed the required statistics and loaded them before serving it to the user. Here are the graphs that are presented there:

1. Overall count of tweets per country.
2. Overall count of tweets per POI.
3. Average sentiment for covid and vaccine by general population
4. Vaccine Hesitancy word cloud
5. Percentage of tweet as per Vaccine/Covid/Neither.
6. Percentage of tweet as per language.

We have precomputed the sentiment separately and uploaded to Solr. The following methods were used to do sentiment analysis for the different languages.

1. English: We use a python library called flair, which is a state-of-the-art NLP library that performs sentiment analysis for a given dataset. It handles Named Entity Recognition at the backend and gives us the sentiment probability for a given tweet.
2. Hindi: We used a keras LSTM model to predict the sentiment of a given tweet. We used a publicly available data for creating the training and testing dataset.
3. Spanish: We used an off-the shelf library called sentiment-analysis-spanish. It readily gave us the sentiments without any need for training.

Once the sentiments were computed, the dataset was transformed to fit the narrative in the analytics dashboard.

For vaccine hesitancy we tried to find keywords that indicates towards vaccine hesitancy and tried to find them in our tweets. If found, we classified those tweets as vaccine hesitancy. Some of those keywords are:

“mybodymychoice”, “NoVaccineForMe”, “NoForcedVaccines”, “stopmandatoryvaccination”, “forcedvaccines”, “covidvaccineispoison”, “VaccinesAreNotTheAnswer”, “medicalfreedomofchoice”, etc. In total there were 55 keywords we used to find the results.

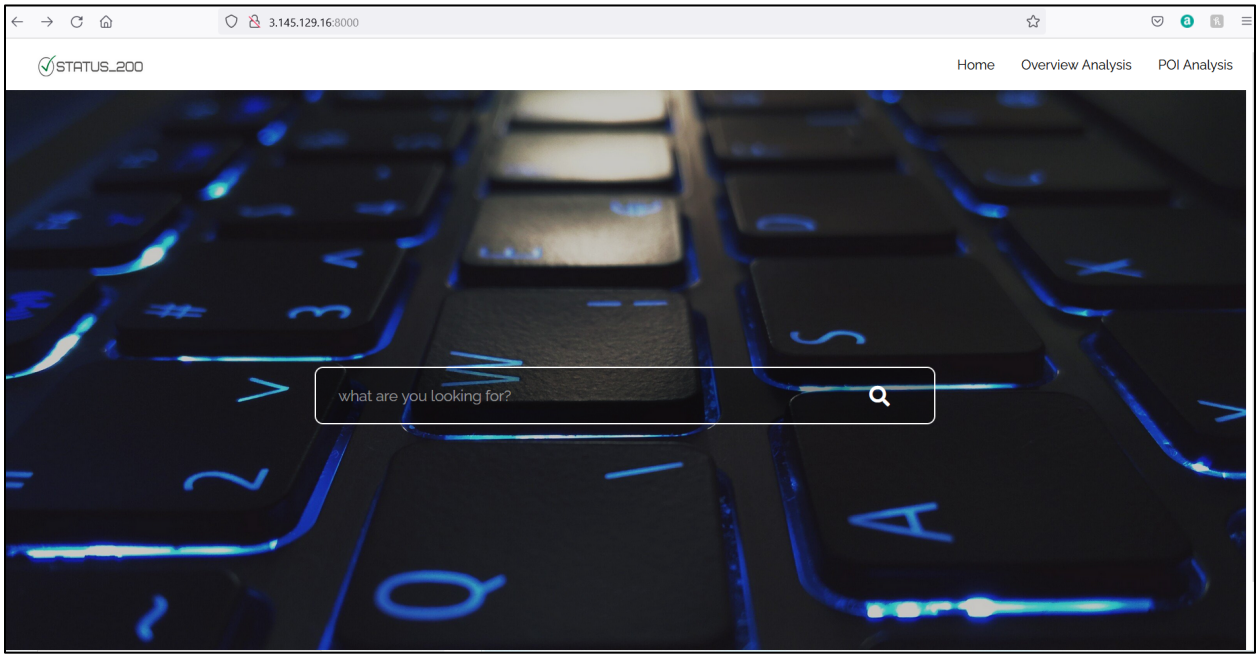
Based on the tweets that were marked true for vaccine hesitancy, we tried to clean the tweet text and create a word cloud of the words that help us understand the most frequent words used in those kinds of tweets.

Finally, we have created another dashboard for POI analysis. Here are the graphs displayed over there (Based on filter on POI names):

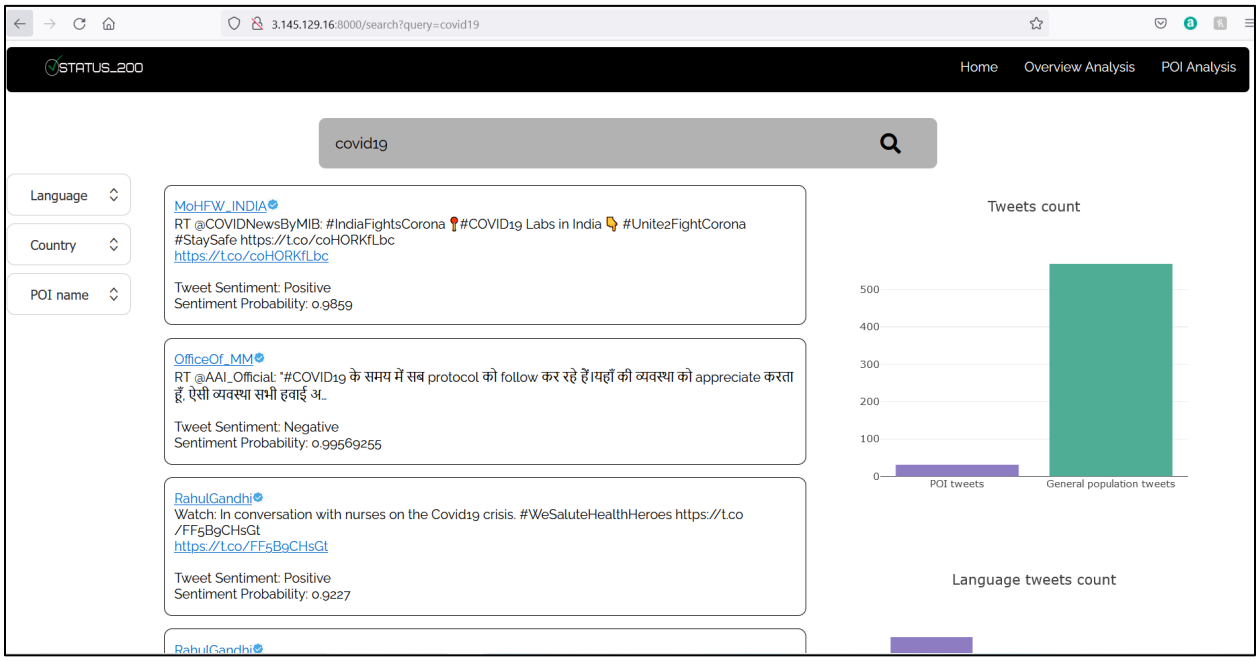
1. Type of tweets count (Vaccine/Covid/Neither)
2. Overall average sentiment for POI.
3. Heatmap for Date-wise number for POI.
4. Curve for new Covid cases based on country, upon which we are displaying instances of covid and vaccine tweets per poi.

Screenshots:

Homepage:

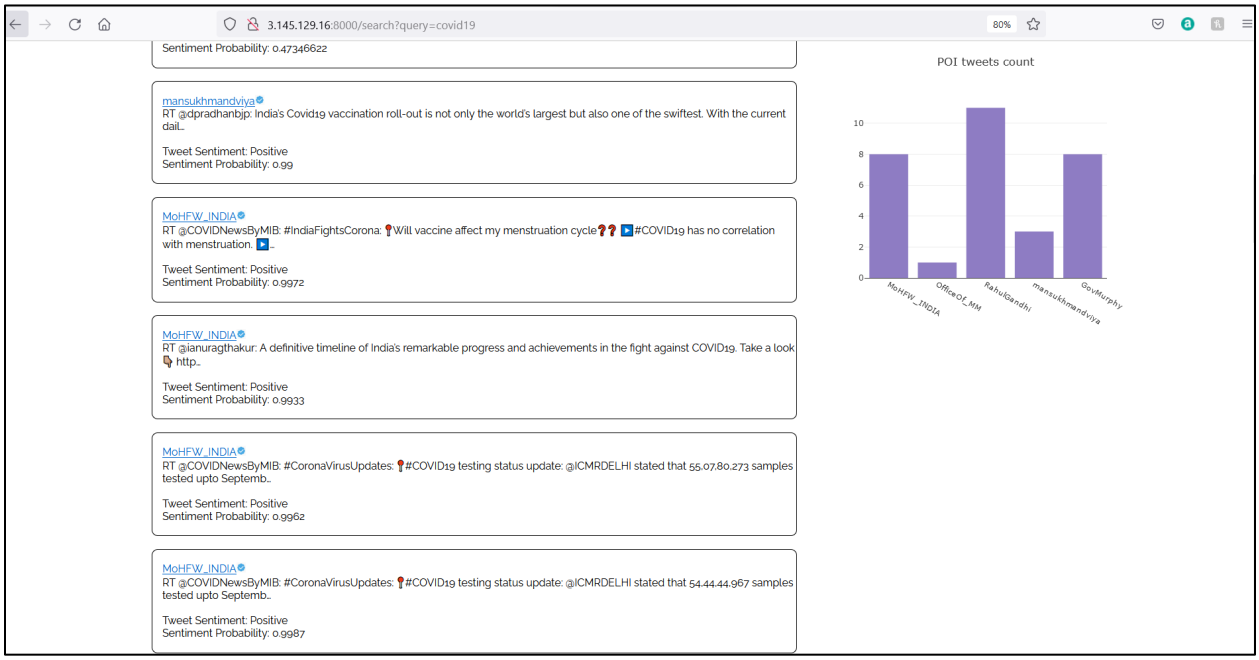


Initial Search Page Results:

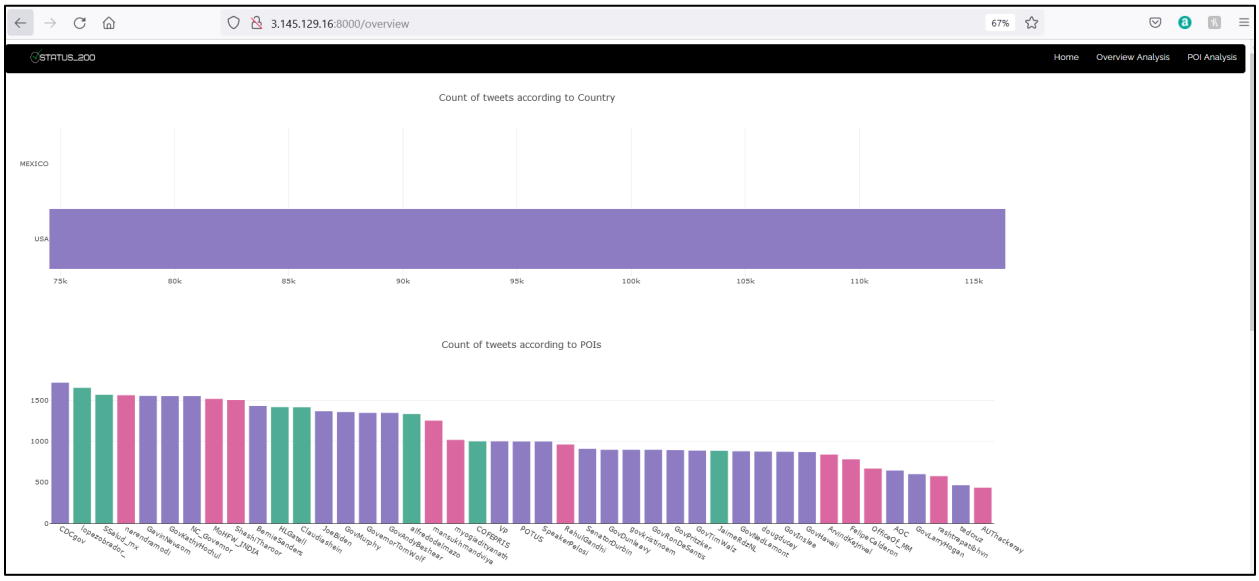


Second Part of Initial Page Results (For the charts on the side):

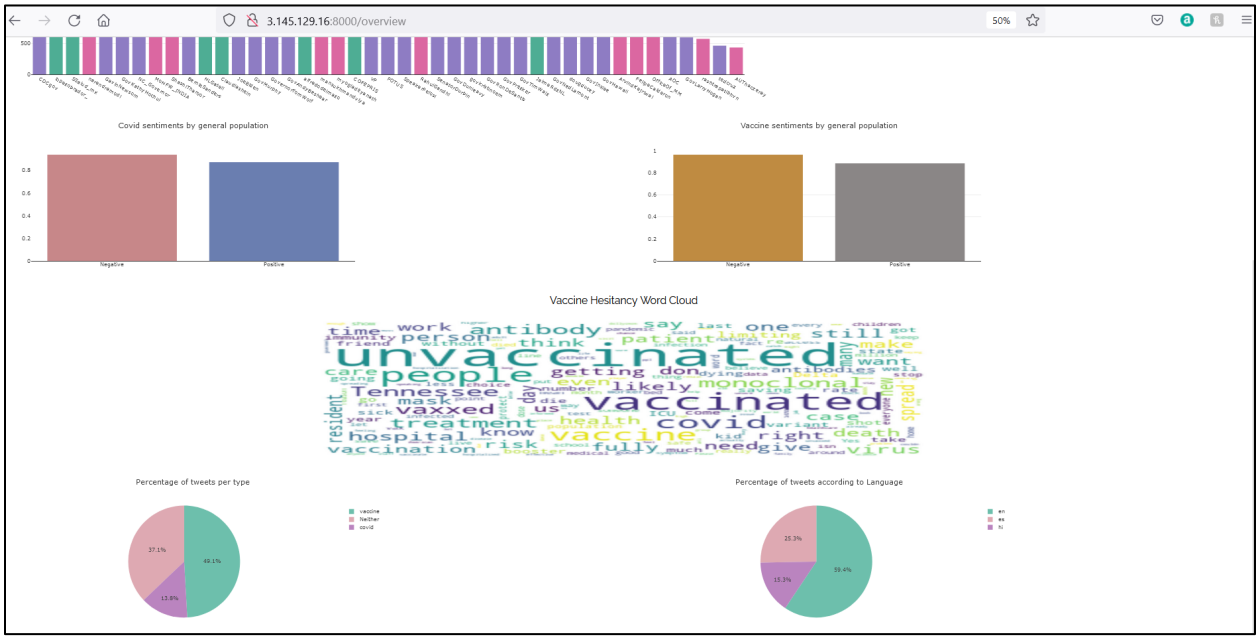
Third Part of the Initial Page Results(For the charts on the side):



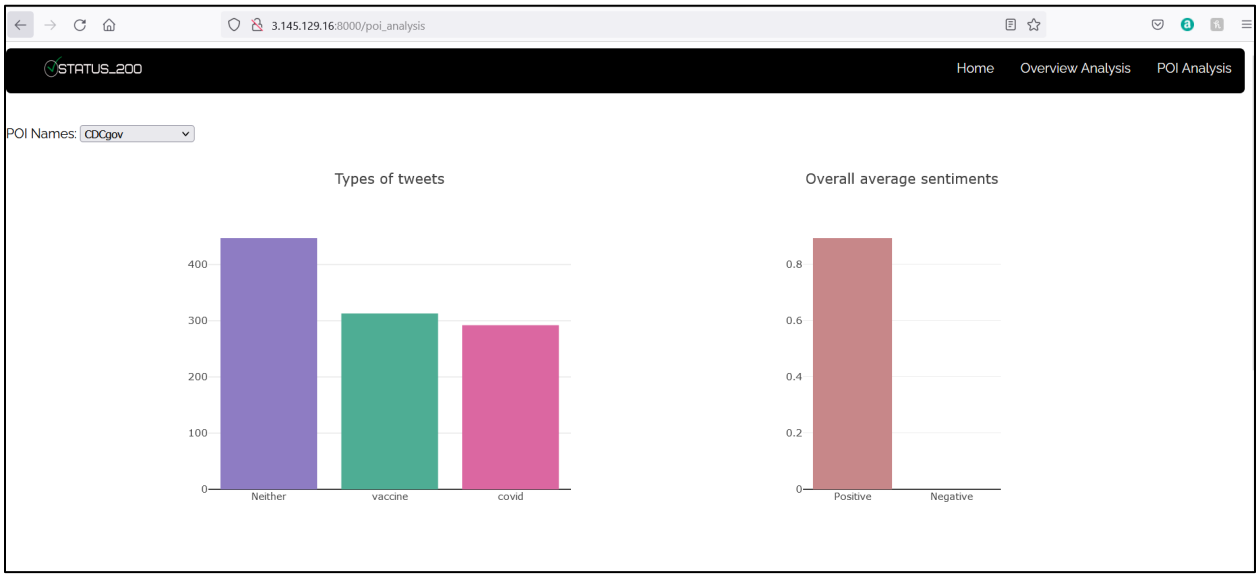
Overview Analysis Part 1:



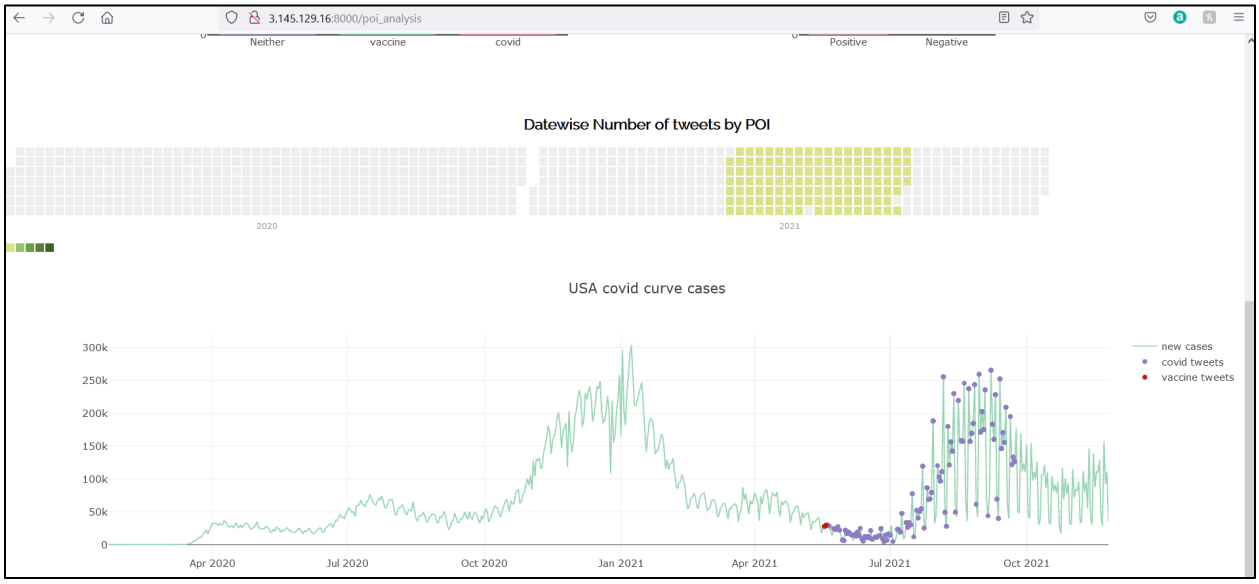
Overview Analysis Part 2:



POI Analysis Part 1:



POI Analysis Part 2:



Team-mate contribution:

Team-mate name	UB IT Name	Responsibilities
Anup Atul Thakkar	anupatul	Creating the backend flask server, handling search query, computing statistics for query results and testing backend for bugs.
Pushkaraj Joshi	pjoshi6	Performing Sentiment Analysis, cleaning and consolidating data-sets, precomputing statistics for overview page, deployment to AWS and user stress testing website for bugs.
Anuja Raghunath Katkar	anujarag	Created the UI of Homepage, query results page, the overview page and POI analysis. Created different types of visualizations on overview and poi analysis page. Also, integrated all of them and handled the overall website functionality
Sagar Jitendra Thacker	sagarjit	Performing topic modelling, exploring and creating graphs offline for story-telling, cleaning and consolidating data-sets and precomputing statistics for overview page.

Conclusion:

This project enabled us to explore and build an end-to-end IR system. Along with that we performed different NLP tasks such as sentiment analysis, and topic modelling. Based on the visualizations we were able to better understand the underlying data and tell a story for the same. Also, we tried to find the correlation between tweets done by POI and the covid curve based on country. Where we could find POI tweeted more about covid than vaccine and, they have been consistent in their number of tweets to communicate to the common masses.