PROJECT 4

CSE 535

INFORMATION RETRIEVAL

By- Latent\_

<https://latent-retrieval.herokuapp.com/>

EXAMINING TWITTER DATA TO ANALYSE GOVERNMENT AND PUBLIC REACTIONS TO COVID

**1.Introduction:**

For analyzing the responses we first collect twitter data based on three countries namely USA, India, Italy and the languages English, Hindi and Italian of the respective countries. In our project we are building an end to end IR system with a search engine and an analytic web UI to present useful insights about COVID 19 and projecting statistical data by examining the collected twitter data. Firstly, we are analyzing the influencer score and using it for ranking tweets. Secondly, we are performing topic analysis of all the tweets to extract the important topics about which people are concerned on the basis of country. Thirdly, we are showing results retrieved from sentimental analysis through visualization with a help of charts, bar graphs and line graphs. The web application also provides additional data like news articles. Lastly, we are showing faceted search operation on the indexed data based on the query, country person of interests and topics.

**2.Dataset:**

The data has been collected from Twitter website based on Twitter API. The total dat set size would be 120k tweets. The data collected is categorized in the following way:

* Tweets based on language namely English, Hindi and Italian.
* Tweets based on country namely India, Italy, USA.
* Tweets based on person of interests of the respective countries.
* Tweets posted in 5 consecutive days focused on reactions of general public to government’s policies on COVID.

**3.Architecture:**

**Requirement 1** : Influencer Score is calculated in this phase of the requirement for every “tweet” collected in the dataset.

The idea to give weightage to every field of twitter based on its value to its corresponding tweet brought us the solution calculate IS based on this formula-

IS = 0.35 \* No. of retweets + 0.1 \* likes + 0.35 \*followers count + 0.1 \*sentiment count + 0.1 \* verified account

Using the above formula, the Influencer Score is calculated for every tweet and are ranked on the basis of that.

**Requirement 2 :**

Topic Analysis is done by considering each and every tweet present in the full\_text field, gensim and LDA models are used to analyze the topics. We have used pyLDAvis to display the analysis in a static HTML page where the user can analyze most frequent topics that are discussed. The algorithm of topic analysis is as follows.

1) Text Cleaning (full\_text) is done by using nltk, removed all the stopwords from all the different languages and tokenized the data.

2) Prepare text for LDA, then add to a list.

3) Create a dictionary from the data, then convert to a bag-of-words corpus and save the dictionary and corpus for future use.

4) We pick the number of topics ahead of time even if we’re not sure what the topics are.

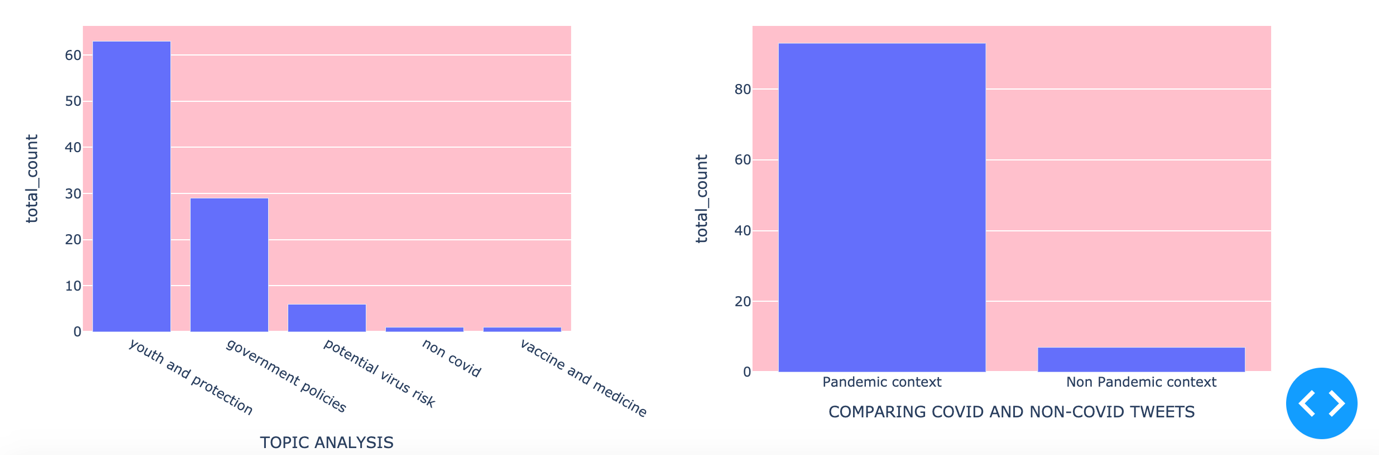
5) Each document is represented as a distribution over topics.

6) Each topic is represented as a distribution over words.

*Correlation analysis:*

Correlation is analyzed between Covid cases in each country, This is achieved by picking up the most important person of the country (President/Prime Minister) and data is normalized since tweets are not keeping up to the increase in the number of covid cases in each country. The data is maintained in a sqlite database and displayed at the point of retrieval.

Interesting facts are obtained from the 5 day period (18-Aug-23-Aug), In India and USA, prime minister and president had made tweets related to covid, as per the number of increase in cases. But in Italy the reverse has happened, when the number of cases are low the tweets related to covid are high and when the cases are high, tweets related to covid are low.



**Requirement 3 :** We have provided insights based on the analysis of above requirements for-

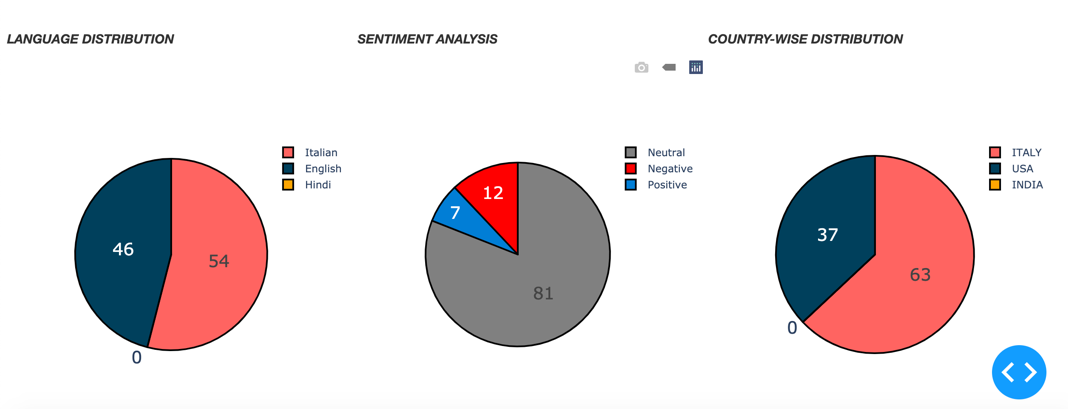
1. Sentiment Analysis: It was done using “vaderSentimentAnalysis” package and a score was obtained for every tweet(during loading). For context in every language was translated to English using “googletranslate” and sentiment score was calculated.

For score > 0.2 , it is labelled as positive

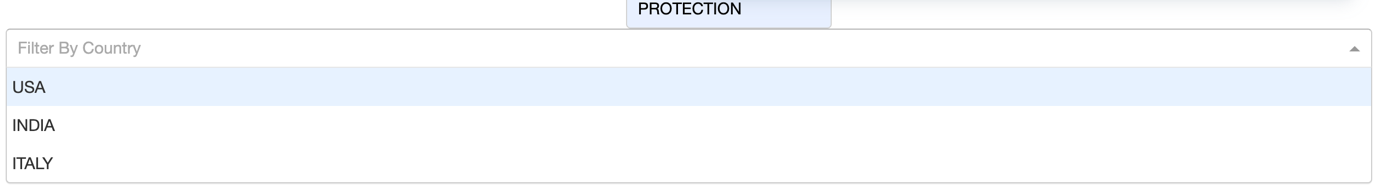
For score < -0.2, it is labelled as negative and

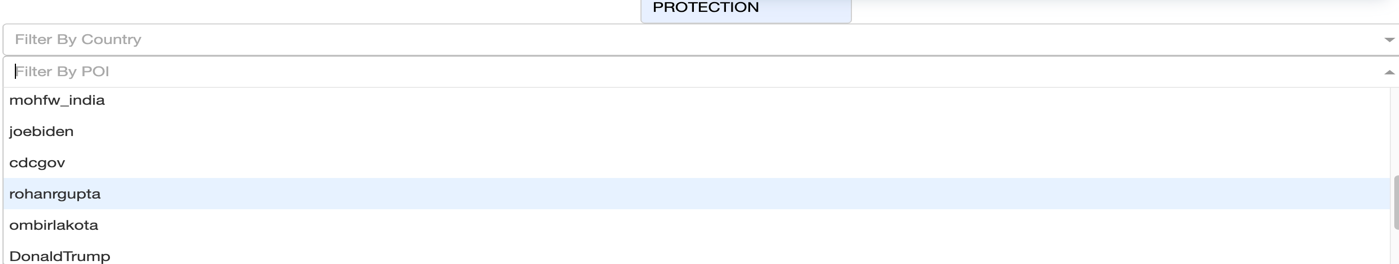
-0.2 to 0.2 as neutral.

1. News Articles: News articles were fetched using “newsapi”, as they are related to search queries with 30 days limit.
2. Visualizations:
3. Language Distribution- A pie chart depicting the share of three languages(English,Hindi and Italian)
4. Sentiment Analysis- A pie chart revealing the contribution of positive,negative and neutral sentiments for the query.
5. Country-wise Distribution- It shows what are the country wise distributions of the tweets loaded.(Italy,USA OR INDIA)
6. Topic Analysis- The bars depicts the counts of tweets related to independent 5 topics as shown in fig.
7. Comparing Covid/Non-Covid Tweets- It shows the amount of pandemic and non-pandemic content retrieved for the query performed.



**Requirement 4:** Multi faceted search: This on choice search was implemented where a user can search for any specific country or any specific POI or any combination of the two to get the desired results of that particular query.





**5) Architecture:**

Front end - HTML, CSS

Back End - Apache SOLR, News API

Framework - Dash, Flask

**6) Conclusion:**

The Information Retrieval project comprised of indexing, analyzing and extracting the data in a way to build end-to-end search engine. We got to understand the concepts of Solr and how much more can be efficiently done with it. We did learn from many new interesting things in the process of getting errors to User Interface and more. I was a whole new idea to understand the data and set those visualizations as what all we can study and get insights from the data through Topic modelling or Geo distribution.

**7) References:**

1. <https://lucene.apache.org/solr/guide/8_5/>
2. <https://newsapi.org/docs>
3. <https://dash-bootstrap-components.opensource.faculty.ai/docs/components/layout/>
4. <https://community.plotly.com/t/setting-the-layout-of-a-button-and-input-box/6519/2>
5. <https://getbootstrap.com/docs/4.0/components/pagination/Vev>

**Contributions:**

Emily Usinski(emilyusi,): Report and collecting and fetching newapi, creating beta interfaces for the web app.

Hitesh Nagothu(hiteshna,50336840): Solr handling, back end, front end, Requirements 3,4

Sonal Shukla(sonalshu,50336648): research work, front end(css), Requirements1,2