CSE 535:Information Retrieval

MULTILINGUAL SEARCH ENGINE

Team: BITS & BYTES

Karthika Jayaprakash - 50317568
Prakhathi Murugesan - 50316846
Thana Shree Jeevanandam - 50320922

User Experience

Search

The user interface consists of a search bar at the top of the page and a set of filters to the left. There are three tabs to display search results, analytics on the result which shows the trend and composition of data and then the news article.

The tweets can be filtered based on

- Date
- Sentiment
- Language
- Verified
- Country
- Topic

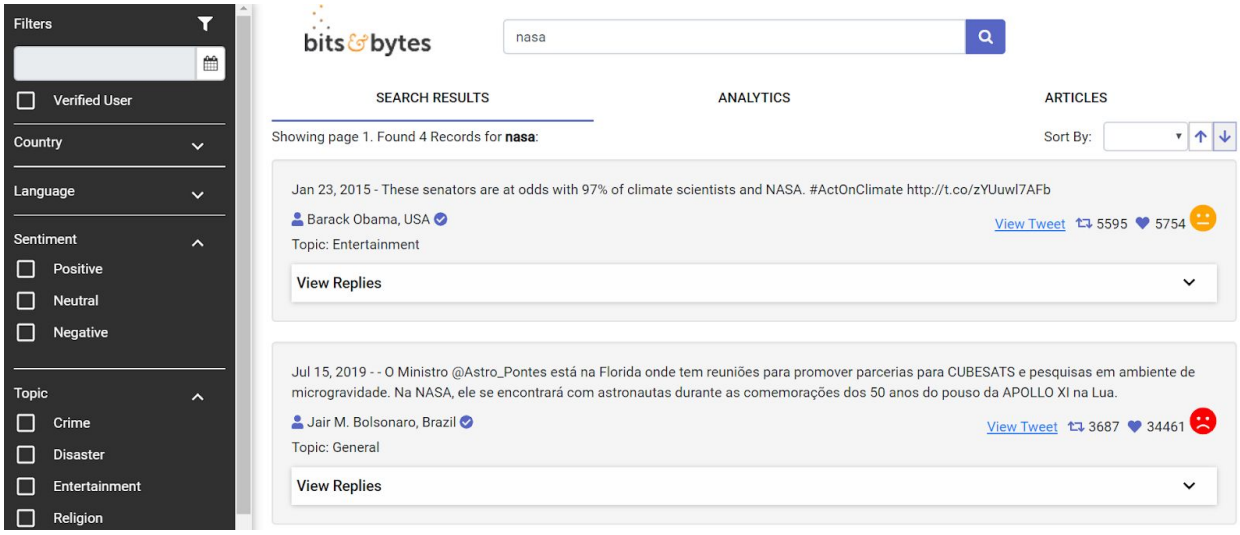
They can be sorted based on

- Solrs' default sort order
- Date
- Likes
- Retweets

The filter and sort parameters are appended on to the solr search URL to get the required tweets. The data can be viewed as individual tweets in the **Search Results** tab or as graphs and charts in the **Analytics** tab. The components are built using ReactJs and Material-UI.

Search Results

The relevant tweets are displayed as a list in the Search Results tab. The results are restricted to top 100 results and 10 results per page are displayed. Pagination is implemented to view the next and previous pages. Information such as tweet text, tweeted date, tweet location, number of likes, number of retweets and the sentiment of a tweet are provided.
A link to the original tweet is provided for the users' convenience.
Each tweet contains a link to see the replies associated with the tweet and sentiment analysis is done on replies.



Analytics

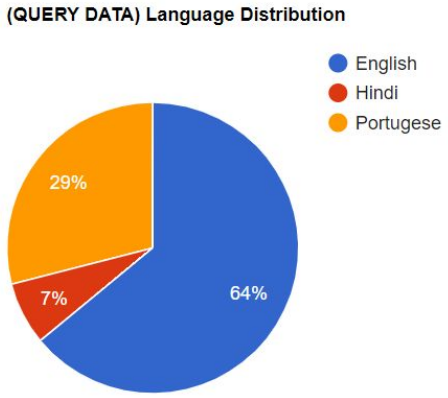
Analytics is done with six variations. The graphs are created with React Google Charts and are integrated using ReactJS. The data to the graphs are taken from the json file retrieved from Solr. Based on the data distribution, the graphs are implemented at two levels,

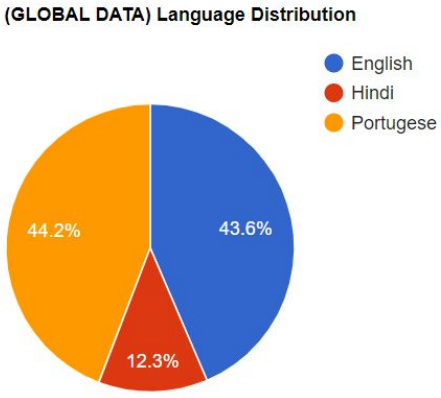
- Query Data - Dynamic graphs representing the analytics of tweets with respect to the particular query searched for.
- Global Data - Static graphs representing the analytics of tweets loaded into the Solr repository.

The analytics tab allows the user to select one among the six different visualizations provided. They include,

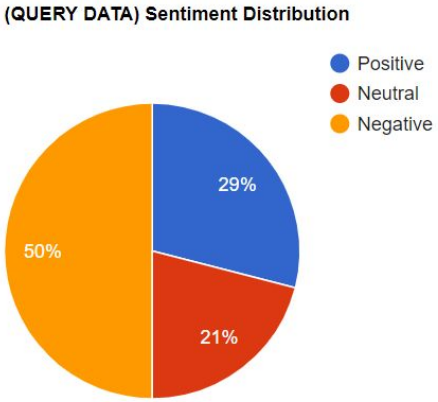
- Language Distribution
- Sentiment Distribution
- Geographical Distribution
- Time Series
- Topic Distribution
- Popularity Analysis

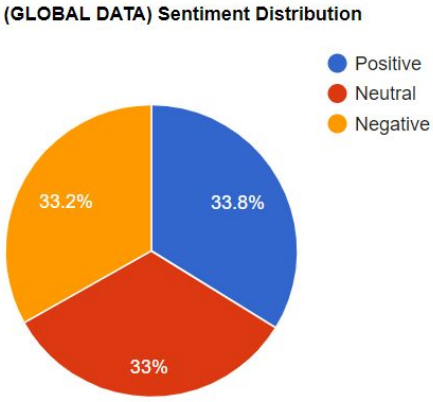
LANGUAGE DISTRIBUTION: The multilingual functionality of the project is analyzed in this distribution. The Pie Chart is plotted with the language and its count of tweets for the particular query. The graph is dynamic and changes as and when different queries are searched for.





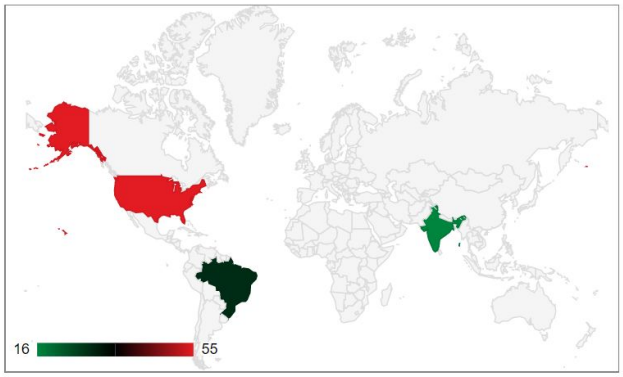
SENTIMENT DISTRIBUTION: The Sentimental Analysis functionality of the project is analyzed in this distribution. The Pie Chart is plotted with the Sentiment of tweet and its count. The graph is dynamic and changes as and when different queries are searched for.



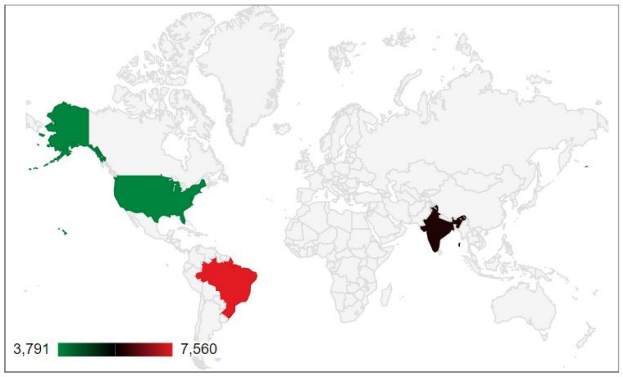


GEOGRAPHICAL DISTRIBUTION: The distribution of the tweets are given with respect to their geography of origin. The red shade gives the country with the highest tweet count for the Query data and the Global data. The graph shows that all tweets share their origin among India, the United States, and Brazil.

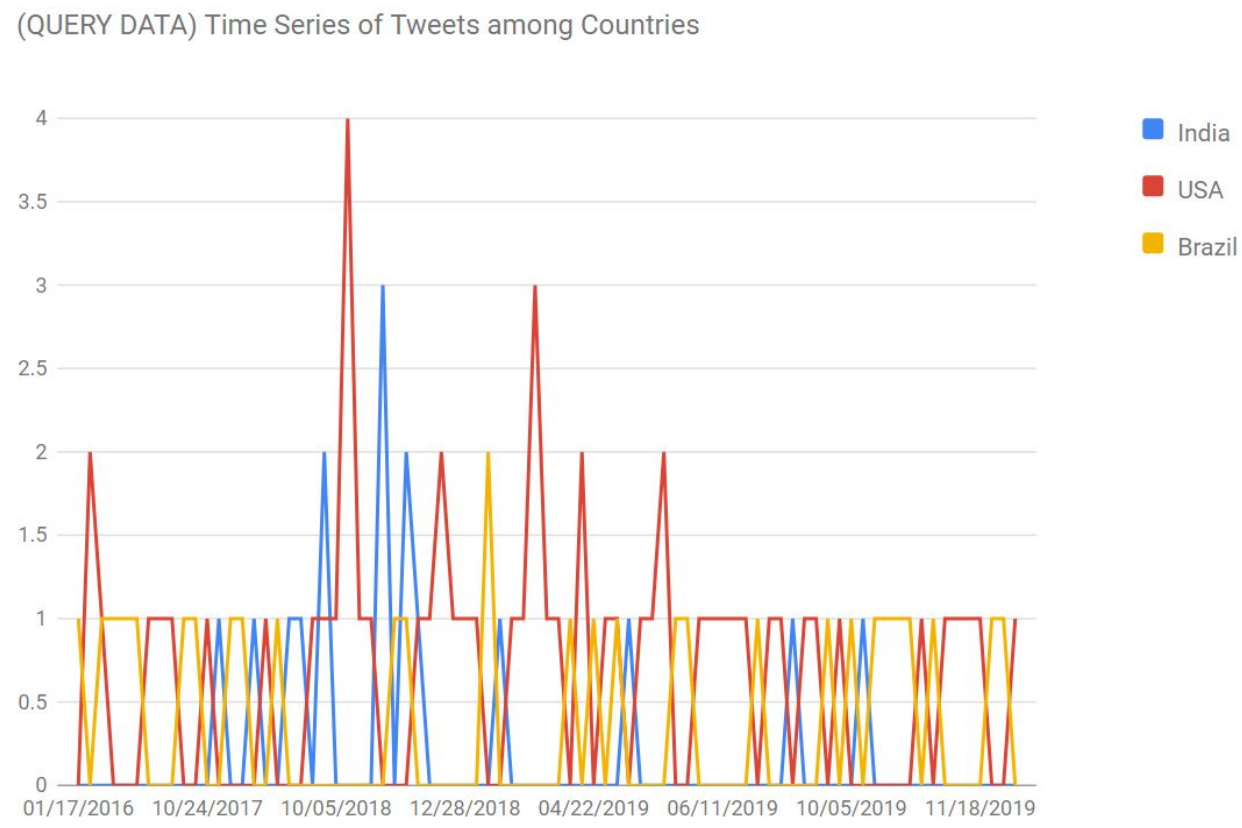
(QUERY DATA) Tweet Distribution across the World



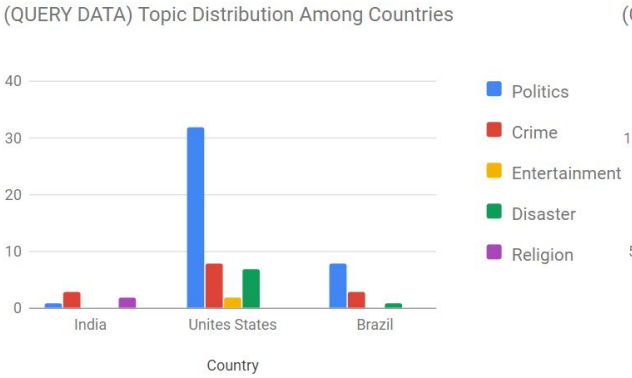
(GLOBAL DATA) Tweet Distribution across the World

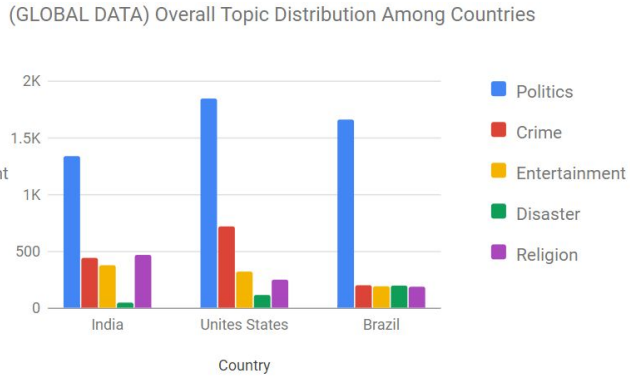


TIME SERIES: The periodicity of the tweets are shown using the time-series graph. The graph plots the count of tweets from India, the United States and Brazil across the date of the tweet posted. With the spike among the graphs, the important dates among the country can be explored.

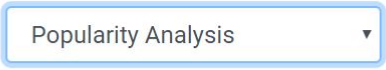


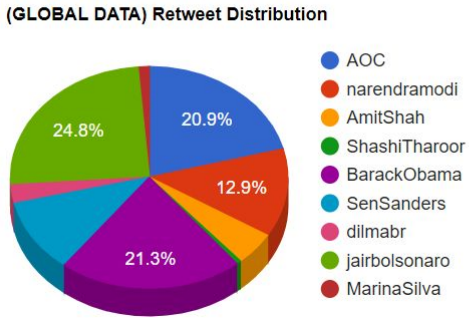
TOPIC DISTRIBUTION: The Topic Modelling functionality is encountered with the plot of tweet count with respect to their topics. As the project deals with Political Analysis, the topic “Politics” seems to top the list in the graph.

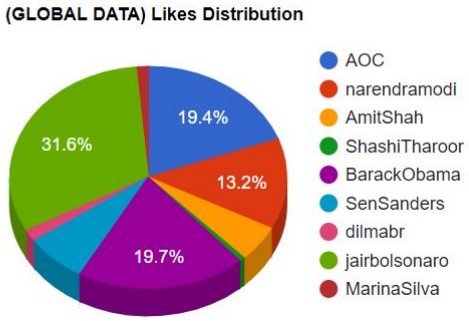




POPULARITY ANALYSIS: The Popularity Analysis analyses the Retweets and Likes Distribution among the various Persons of Interest chosen. It gives the ranking of the Persons of Interest with respect to their retweeted and total likes count for the tweets stored in Solr repository.

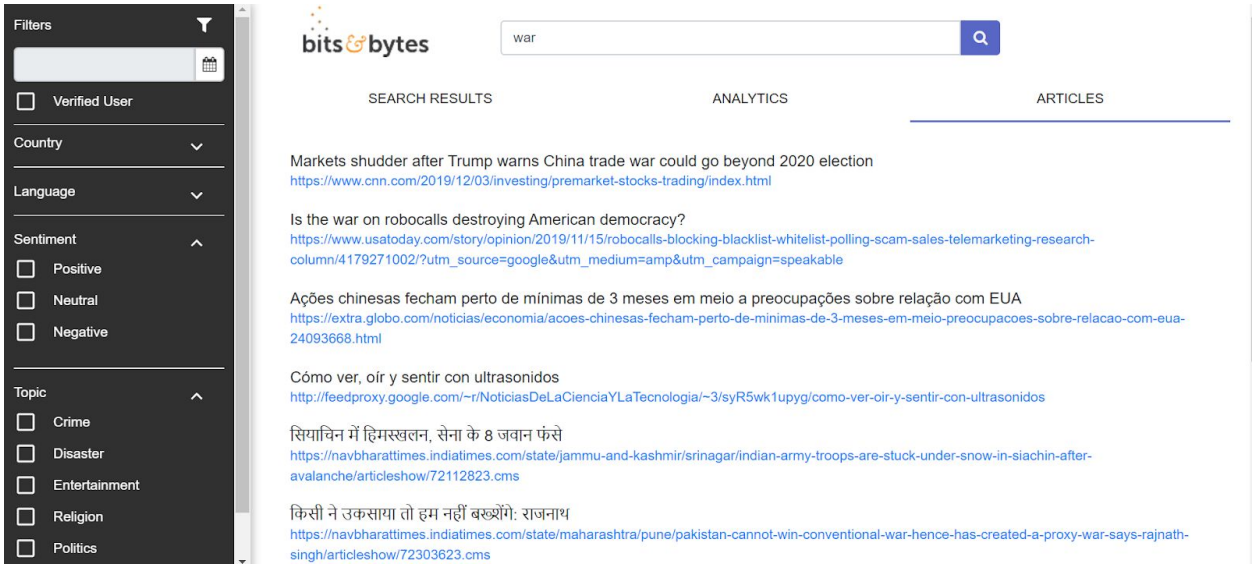






Article

The news articles are retrieved for the particular query using NewsAPI which provides articles in different languages. Each article section consists of the title and a link to view the article.
The NewsAPI querying is similar to Solr's query functionality. We can specify the query words as well as a filter on countries, languages and news sources.
In order to get a balanced retrieval of articles, we have specified the sources so that articles of all languages (Hindi, Portuguese and English) and all topics are retrieved. The articles also support multilingual querying.

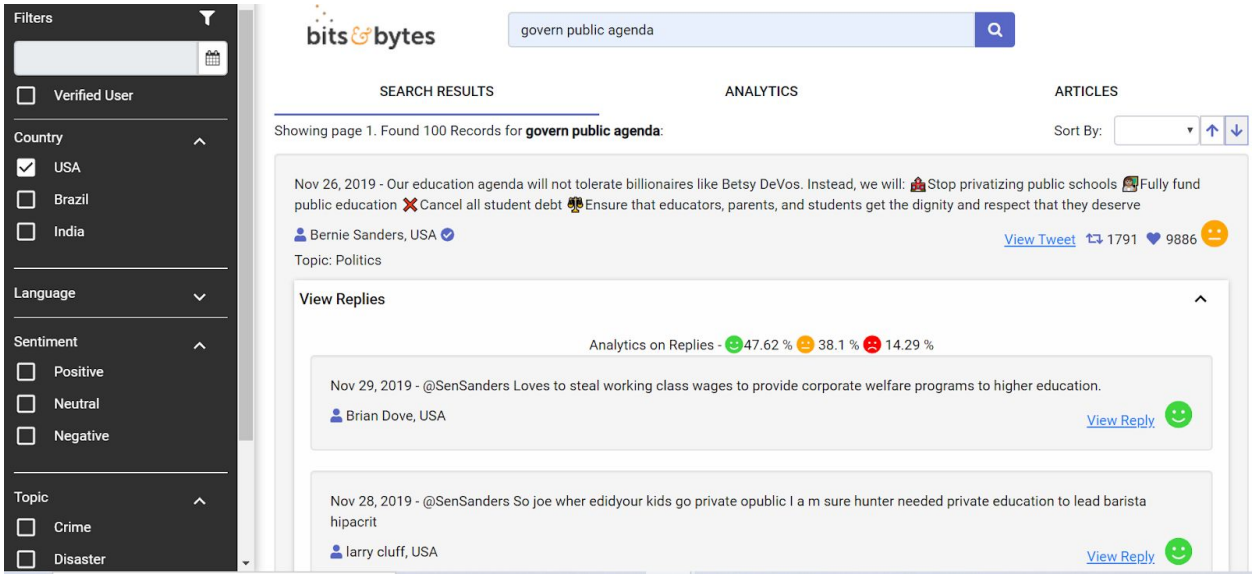


Data Analytics

Sentiment Analysis

Sentiment Analysis of tweets is computed on two levels.

- Each tweet text is analyzed for its sentiment.
- The replies to a particular tweet are analyzed and the overall sentiment spread is calculated.



The initial approach involved the use of the googletrans and TextBlob library. The Spanish and English text is translated to English and then sentiment analysis is performed. However, the API hit limit restricted the usage on all tweets.
The next approach was to build a sentiment analyzer based on the AFINN-111 dataset and rtatman dataset which score words positively or negatively based on their sentiment.
The tweet is tokenized and scored using the model to obtain the sentiment associated with it.

Topic Modelling

The frequently occurring words were grouped using LDA and the clusters were manually assigned the labels Crime, Politics, Religion, Disaster, and Entertainment.
Each tweet was assigned to one of these topics depending on the presence of certain words. We observed that the distribution of the tweets was skewed in terms of Politics but uniform in terms of other topics.

Contributors

Prakhathi Murugesan	User Interface, Data Collection, and Preprocessing
Karthika Jayaprakash	Sentiment Analysis and Topic Modelling on tweets
Thana Shree Jeevanandam	Analytics on Google Charts and News article Retrieval