CSE 535 Information Retrieval – Project 4 Report

Multilingual Search Engine for Analyzing Twitter data

Team: IR Cluster
Department of Computer Science
University at Buffalo, NY 14214

Overview:

The aim of this project is to implement an end-to-end IR solution including a search website which helps the user to perform search operation and return the meaningful results. Along with the search results, various Analytics and Visualizations are shown on the page to the user and it provides a meaningful insight over the collected POIs' tweets dataset and its corresponding impact on society/country.
The end-to-end IR based website is hosted on AWS and available to perform search operation.

Datasets:

1) POI's tweets including retweets and trending hashtags collected over a period of time from three different countries namely USA, India and Brazil in three different languages namely English, Hindi and Portuguese.

Implementation:

1) The tweets dataset collected are in the json file format and has been formatted to use by Elasticsearch.
2) Elasticsearch has been used as search engine for indexing the dataset and performing search operation.
3) Kibana dashboard has been used to showcase the various types of analytics and visualization on the indexed dataset.
4) An end-to-end IR website has been hosted on AWS for end-user as search-engine tool to perform search operation and gets a visualization on the available dataset.

Tools used:

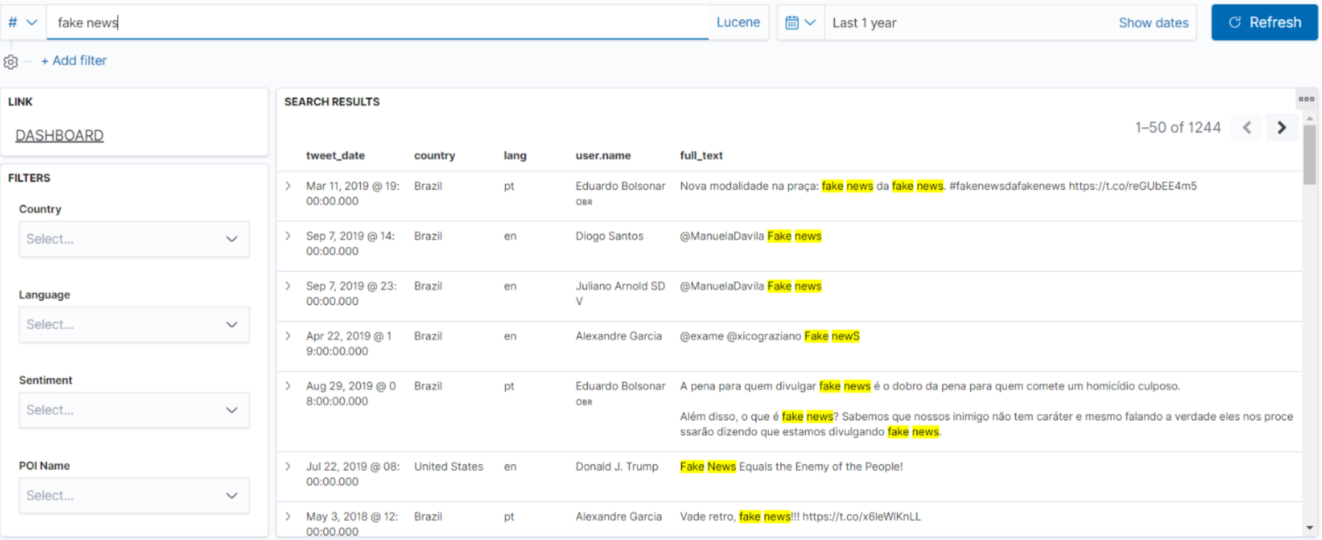
Elasticsearch, Kibana Dashboard, Vader Sentiment analysis tool, Google cloud translate API, Python scripts, AWS

Website tabs:

1) **Analytics:** includes search results and visualizations on complex query searched by the user and provides an in-depth analysis and sentiment on tweets from different POIs.
2) **Dashboard:** includes analytics and visualizations on entire dataset and provides a realistic way of analyzing the impact of tweets over the society and country

CSE 535 Information Retrieval – Project 4 Report

Search Operations and the results:



fake news

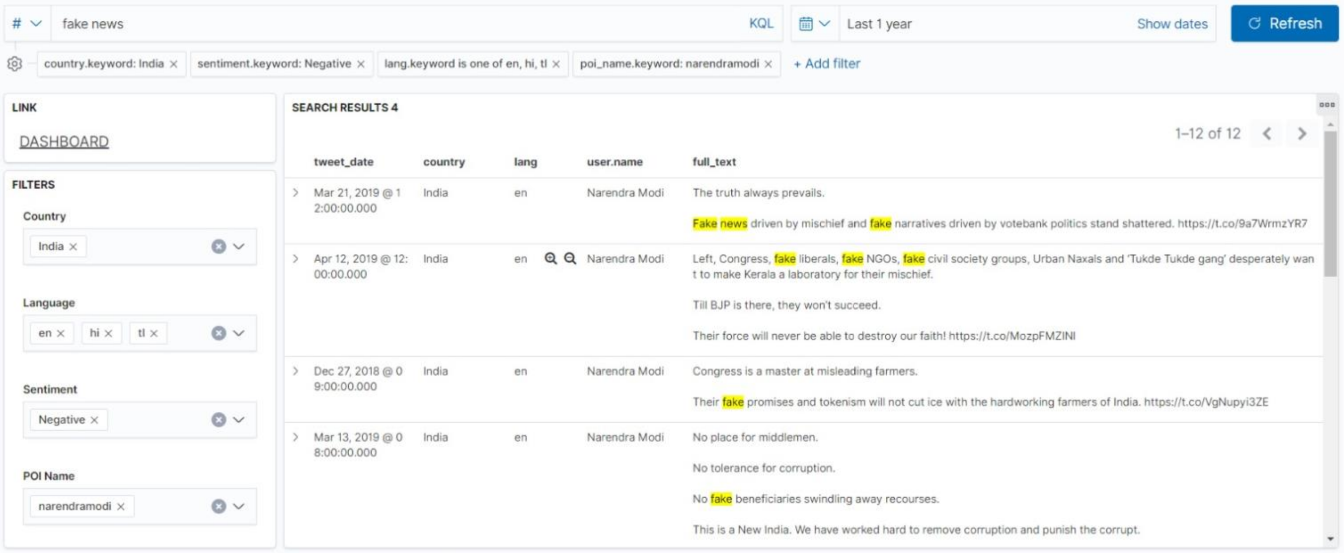
SEARCH RESULTS

1–50 of 1244

tweet_date	country	lang	user.name	full_text
Mar 11, 2019 @ 19:00:00.000	Brazil	pt	Eduardo Bolsonar OBR	Nova modalidade na praça: fake news da fake news. #fakenewsdafakenews https://t.co/reGUbEE4m5
Sep 7, 2019 @ 14:00:00.000	Brazil	en	Diogo Santos	@ManuelaDavila Fake news
Sep 7, 2019 @ 23:00:00.000	Brazil	en	Juliano Arnold SDV	@ManuelaDavila Fake news
Apr 22, 2019 @ 19:00:00.000	Brazil	en	Alexandre Garcia	@exame @xicograziano Fake newS
Aug 29, 2019 @ 08:00:00.000	Brazil	pt	Eduardo Bolsonar OBR	A pena para quem divulgar fake news é o dobro da pena para quem comete um homicídio culposo. Além disso, o que é fake news? Sabemos que nossos inimigo não tem caráter e mesmo falando a verdade eles nos processarão dizendo que estamos divulgando fake news.
Jul 22, 2019 @ 08:00:00.000	United States	en	Donald J. Trump	Fake News Equals the Enemy of the People!
May 3, 2018 @ 12:00:00.000	Brazil	pt	Alexandre Garcia	Vade retro, fake news!!! https://t.co/x6leWIKnLL

Faceted Search:

There are different facet search options have been included on the left side of the page and it makes the website more user friendly. It helps the user to filter out and narrow down the search results based on specific individual filter criteria or set of multi selected filter criteria.



fake news

country.keyword: India × | sentiment.keyword: Negative × | lang.keyword is one of en, hi, tl × | poi_name.keyword: narendramodi ×

Country: India ×

Language: en × hi × tl ×

Sentiment: Negative ×

POI Name: narendramodi ×

SEARCH RESULTS 4

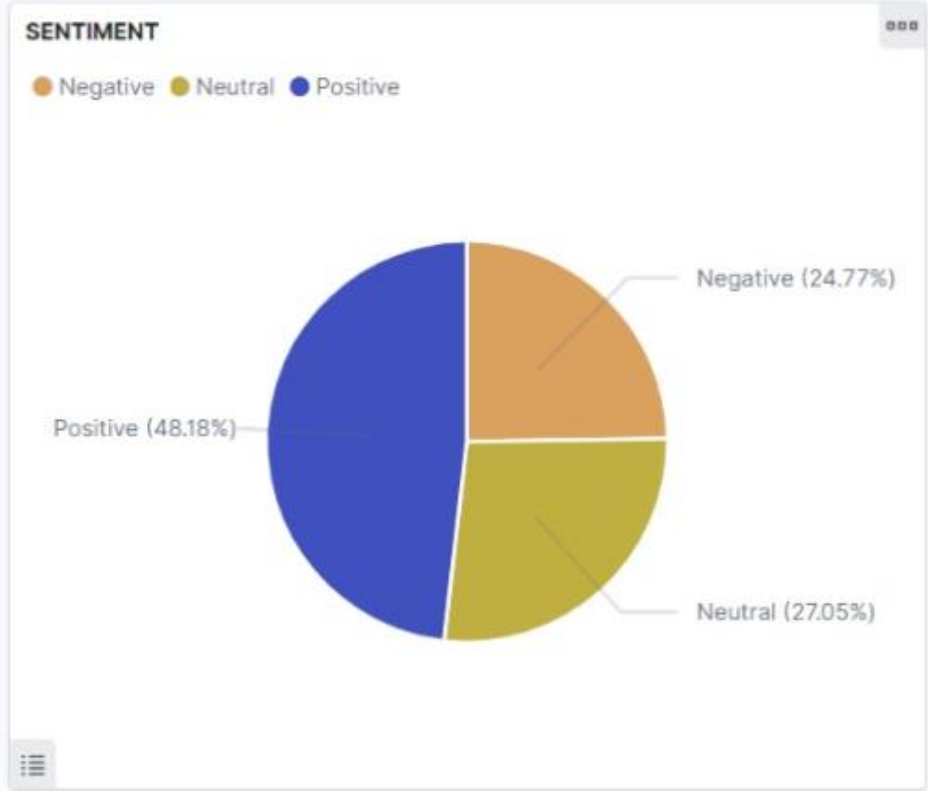
1–12 of 12

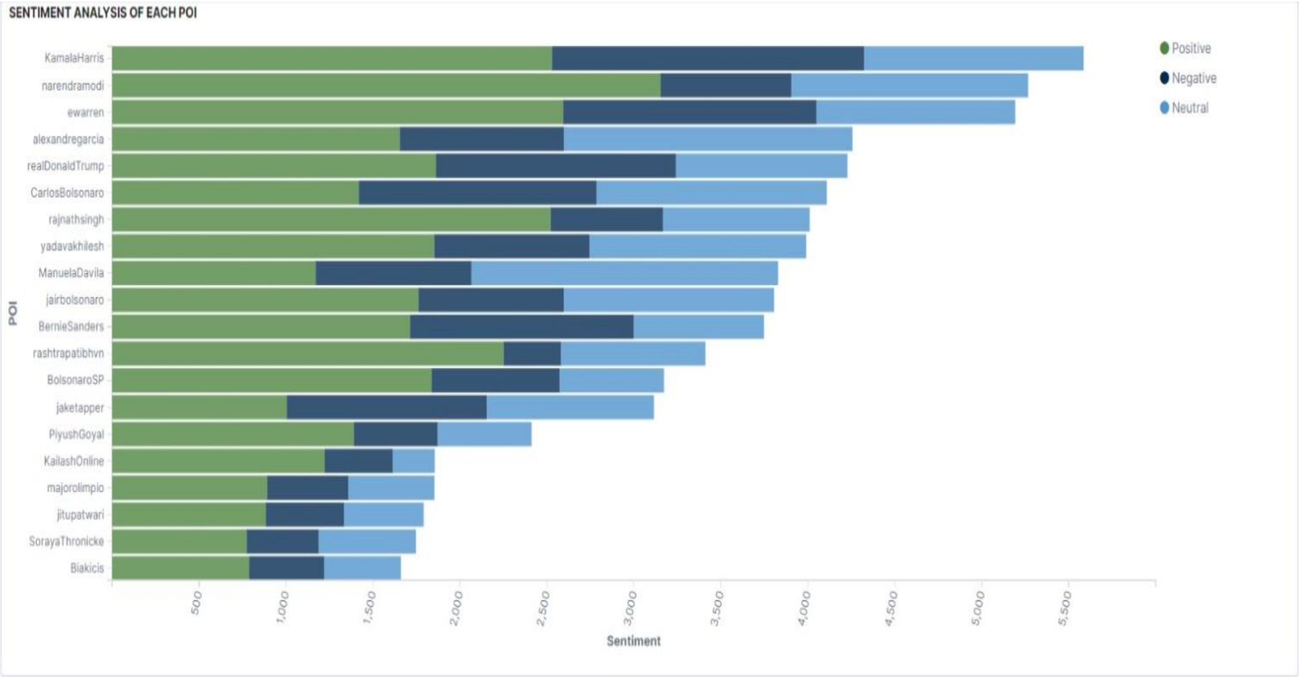
tweet_date	country	lang	user.name	full_text
Mar 21, 2019 @ 12:00:00.000	India	en	Narendra Modi	The truth always prevails. Fake news driven by mischief and fake narratives driven by votebank politics stand shattered. https://t.co/9a7WrmzYR7
Apr 12, 2019 @ 12:00:00.000	India	en	Narendra Modi	Left, Congress, fake liberals, fake NGOs, fake civil society groups, Urban Naxals and 'Tukde Tukde gang' desperately want to make Kerala a laboratory for their mischief. Till BJP is there, they won't succeed. Their force will never be able to destroy our faith! https://t.co/MozpFMZINI
Dec 27, 2018 @ 09:00:00.000	India	en	Narendra Modi	Congress is a master at misleading farmers. Their fake promises and tokenism will not cut ice with the hardworking farmers of India. https://t.co/VgNupyi3ZE
Mar 13, 2019 @ 08:00:00.000	India	en	Narendra Modi	No place for middlemen. No tolerance for corruption. No fake beneficiaries swindling away recourses. This is a New India. We have worked hard to remove corruption and punish the corrupt.

Analytics and visualizations:

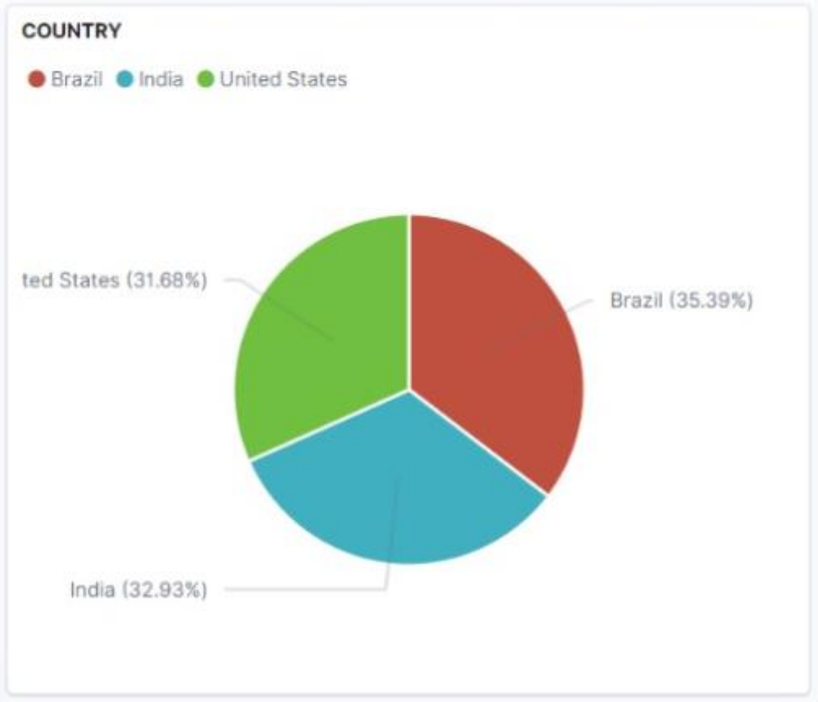
1) **Sentiment Analysis:** Sentiment analysis was performed on each of the tweets using vaderSentiment tool. Before performing the sentiment analysis and getting the sentiment score, we used google cloud API to translate the non-English tweets and applying sentiment analysis on those translated tweets. It helps the user to analyze the tweets sentiment and deep dive through to know how and what impact the tweets and the replies have over the

society whether people are more positive about a tweet/topic or more negative or they have the neutral reaction.

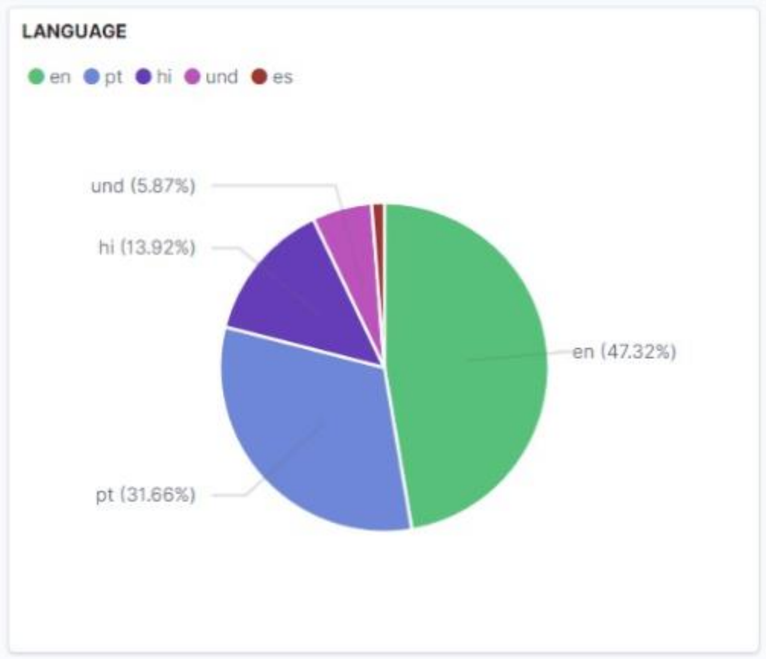




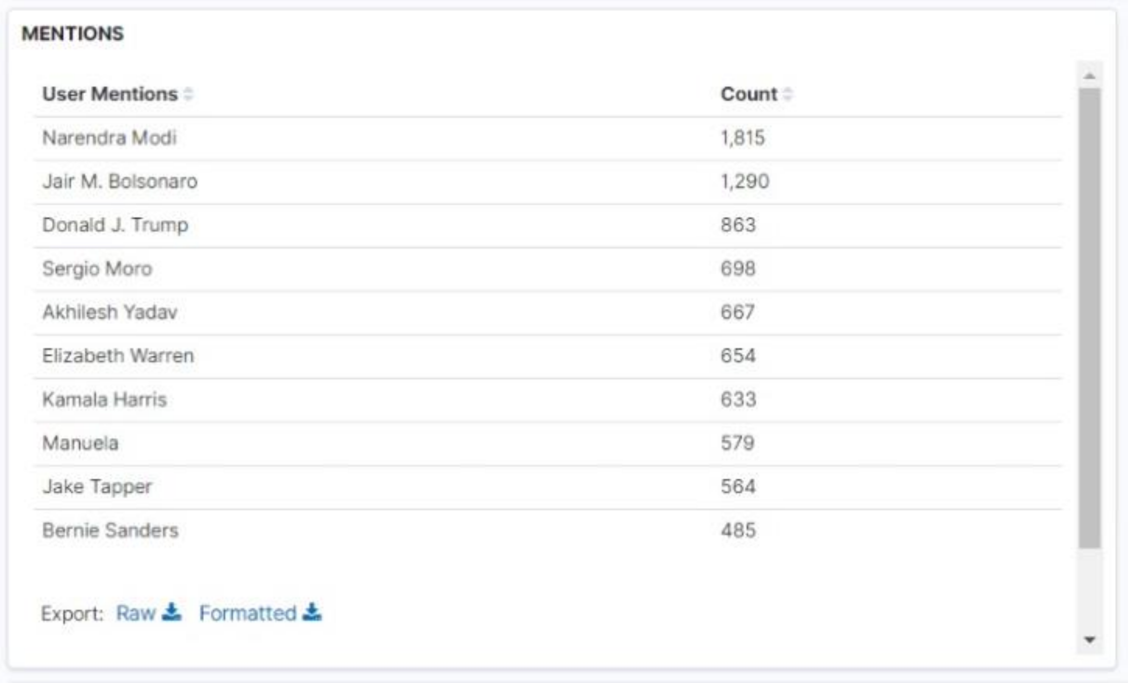
2) **Number of tweets count percentage from each country:** This Kibana visualization gives the number of tweets percentage from each country based on the searched query.



3) **Language:** This pie chart shows the language distribution of the tweets returned from the query result. The user can explicitly get an idea that people are tweeting in which language more on a particular topic or query.



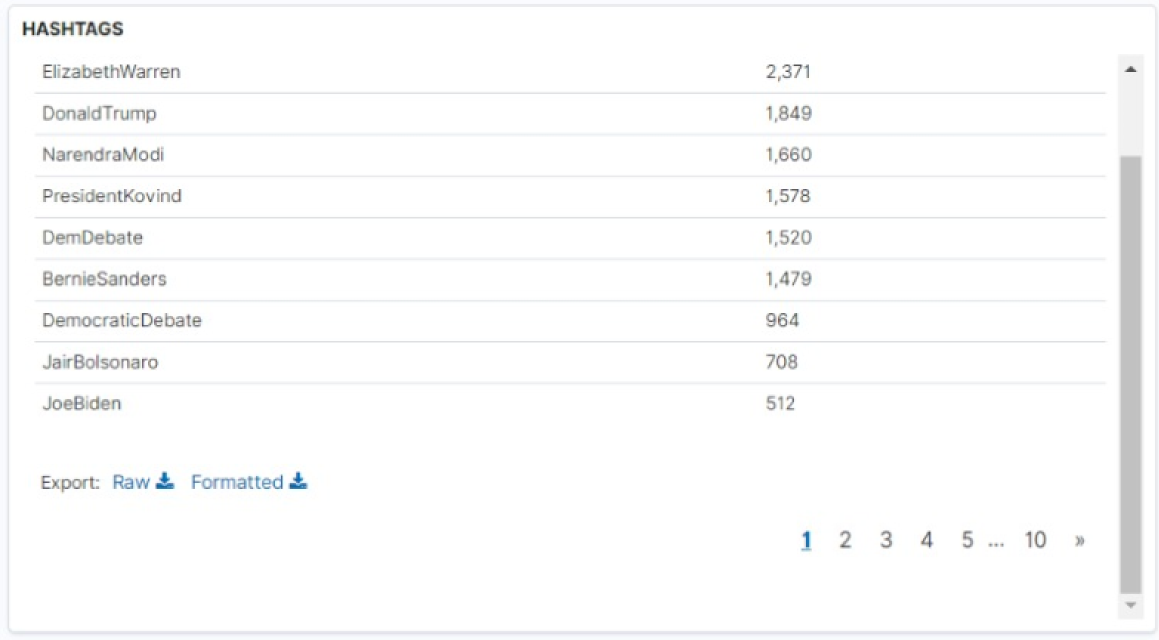
4) **Mentions:** This contains the top mentions which may consists of any famous personality mention or entities like news channel etc.



MENTIONS

User Mentions	Count
Narendra Modi	1,815
Jair M. Bolsonaro	1,290
Donald J. Trump	863
Sergio Moro	698
Akhilesh Yadav	667
Elizabeth Warren	654
Kamala Harris	633
Manuela	579
Jake Tapper	564
Bernie Sanders	485

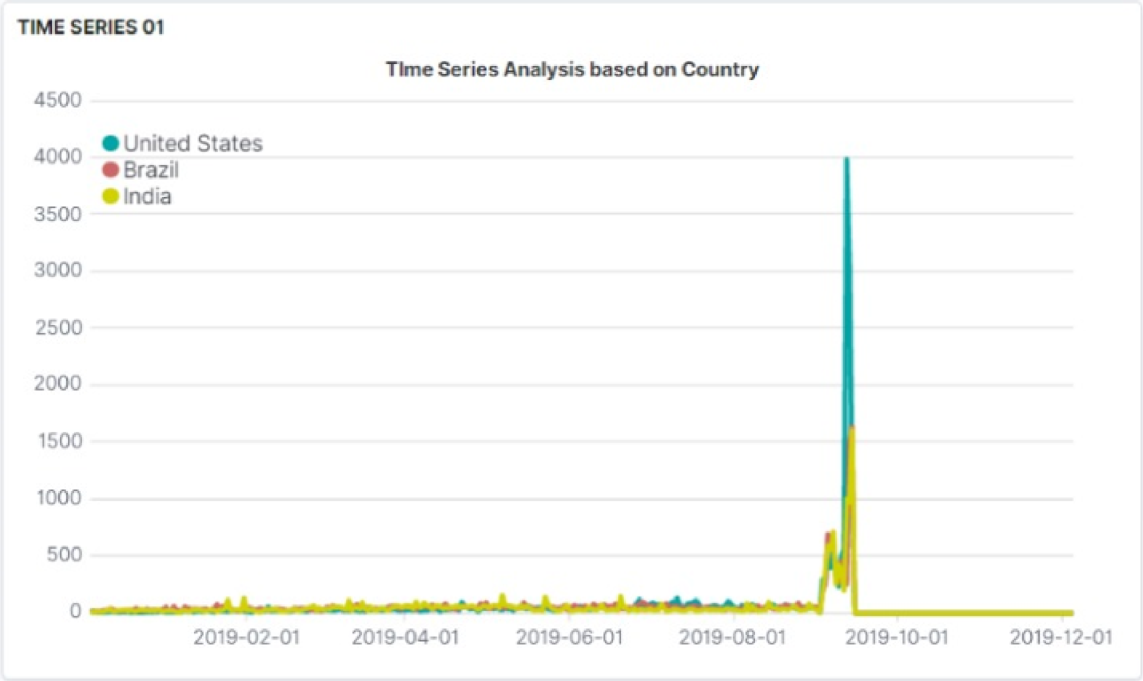
5) **Top Hashtags:** These analytics gives an idea of the top trending hashtags used in the tweets. This is implemented on both dynamic data and on entire dataset and gives an idea of trending topics or person on the twitter.



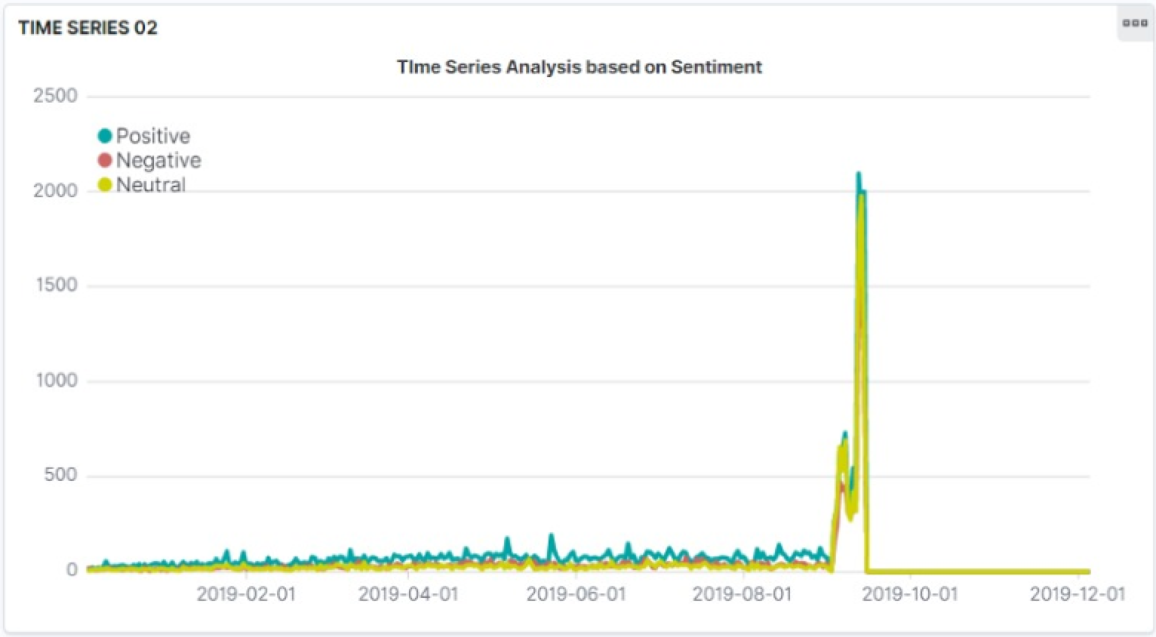
HASHTAGS

ElizabethWarren	2,371
DonaldTrump	1,849
NarendraModi	1,660
PresidentKovind	1,578
DemDebate	1,520
BernieSanders	1,479
DemocraticDebate	964
JairBolsonaro	708
JoeBiden	512

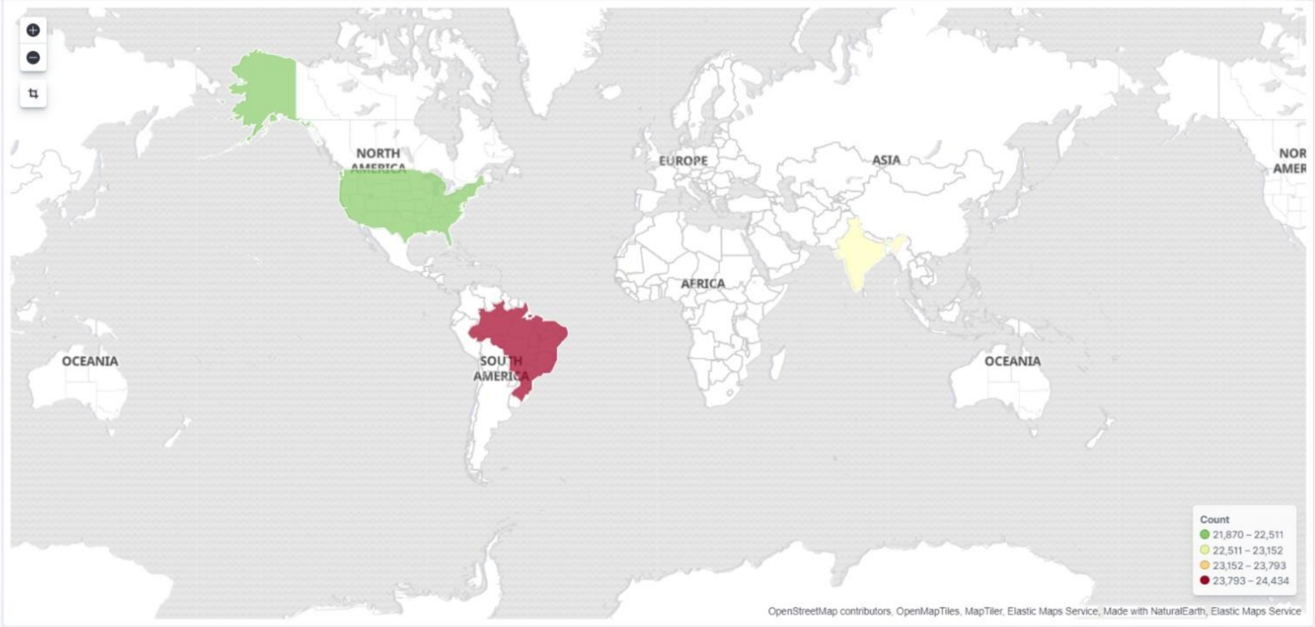
6) **Time Series Analysis based on Country:** This visualization gives the idea that number of tweets generated from each country during different time interval say per day. User also has the option to see time series analysis for individual country after disabling the other countries option by clicking on the legends option including all other graphs.



7) **Time Series Analysis based on Sentiment:** This graph gives a visual insights of tweets sentiment generated on per day basis and helps the user to analyze how many tweets are positive, negative or neutral on the particular day from each country location when combinedly viewed.



8) **Region Map based on number of tweets:** This map gives a visual insight of number of tweets from each country and tells us which country has the largest and the least number of tweets.



Conclusion:

An end-to-end IR website hosted on AWS gives ability to the end-user to perform search on collected tweets data and see the analytics and visualizations on the dataset. It helps the user to analyze the various impacts that the POI's data have over the region and country. The website provides a detail analysis on top trending hashtags, the sentiment of each tweets, time-series analysis by country/sentiment, number of tweets per country/language. These analyses in turn help the user to see and analyze the societal impact of these dataset and deep dive into the details on how the POIs tweets impact a particular region/country.

Members contribution:

1) Gowtham Vuppala (gowthamv)
   Visualization, Analysis and Data

2) Gulfam Hussain (gulfamhu)
   Data Scraping and formatting

3) Nikhil Yadav (nyadav2)
   Visualization, Analysis and Data

4) Swaminaathan Pudukottai Jayarman Murali (swaminaa)
   Elastic Search, Sentiment Analysis, Indexing and Data

References:

1. Project 4 description document
2. Elasticsearch
3. Kibana