CSE535 Project 4: Analysis and Visualization of Twitter Data, HAMπ

Abstract

We aim to analyse the impact of political rhetoric globally and on social media. In our project we are extracting data from Twitter and performing analysis of the collected data based on the person of interest, country, language and time period. We have gathered around 200k tweets which mainly consists of tweets made by impactful people (persons of interest) each having at least 1000 tweets. These persons of interest belong to 3 different countries namely India, United States of America and Brazil. Our collected Twitter data is country specific and is mainly comprised of Hindi, English and Portuguese tweets. Our dataset also consists of replies to these tweets which we use to analyse the impact of a tweet by doing sentiment and volume analysis. We have displayed most discussed topics from our dataset by means of topic categorization. We have provided keyword search to the user to get tweets and perform keyword-based analysis. User can also perform a faceted search which is more specific and allows us to get filtered data, based on POI name, languages, date, etc.

1. Introduction:

Twitter is an online microblogging tool that disseminates more than 400 million messages per day, including vast amounts of information about almost all industries from entertainment to sports, health to business etc. One of the best things about Twitter—indeed, perhaps its greatest appeal—is in its accessibility. It's easy to use both for sharing information and for collecting it. Twitter provides unprecedented access to our lawmakers and to our celebrities, as well as to news as it's happening. Twitter represents an important data source for the business models of huge companies as well.

All the above characteristics make twitter a best place to collect real time and latest data to analyse and do any sought of research for real life situations. Thus, we have collected Twitter data to perform Sentiment analysis and Topic analysis to get the impact such as sentiment analysis, keyword analysis, topic analysis, etc.

1.1 Sentiment Analysis

This project addresses the problem of sentiment analysis in twitter; that is classifying tweets according to the sentiment expressed in them: positive, negative or neutral. Twitter is an online micro-blogging and social-networking platform which allows users to write short status updates of maximum length 140 characters. It is a rapidly expanding service with over 200 million registered users - out of which 100 million are active users and half of them log on twitter on a daily basis - generating nearly 250 million tweets per day. Due to this large amount of usage we hope to achieve a reflection of public sentiment by analysing the sentiments expressed in the tweets. Analysing the public sentiment is important for many applications such as firms trying to find out the response of their products in the market, predicting political elections and predicting socioeconomic phenomena like stock exchange. The aim of this project is to develop a functional classifier for accurate and automatic sentiment classification of an unknown tweet stream.

1.2 Topic Analysis

Topic Models are a type of statistical language models used for uncovering hidden structure in a collection of texts. By doing topic modeling, we build clusters of words rather than clusters of texts. A text is thus a mixture of all the topics, each having a specific weight. Tagging, abstract "topics" that occur in a collection of documents that best represents the information in them. There are several existing algorithms you can use to perform the topic modeling. We are using Latent Dirichlet Allocation (LDA) model to extract important topics from our dataset. The below figure gives us a rough estimate of working of LDA model using which we retrieve main topics from our Twitter data collection.

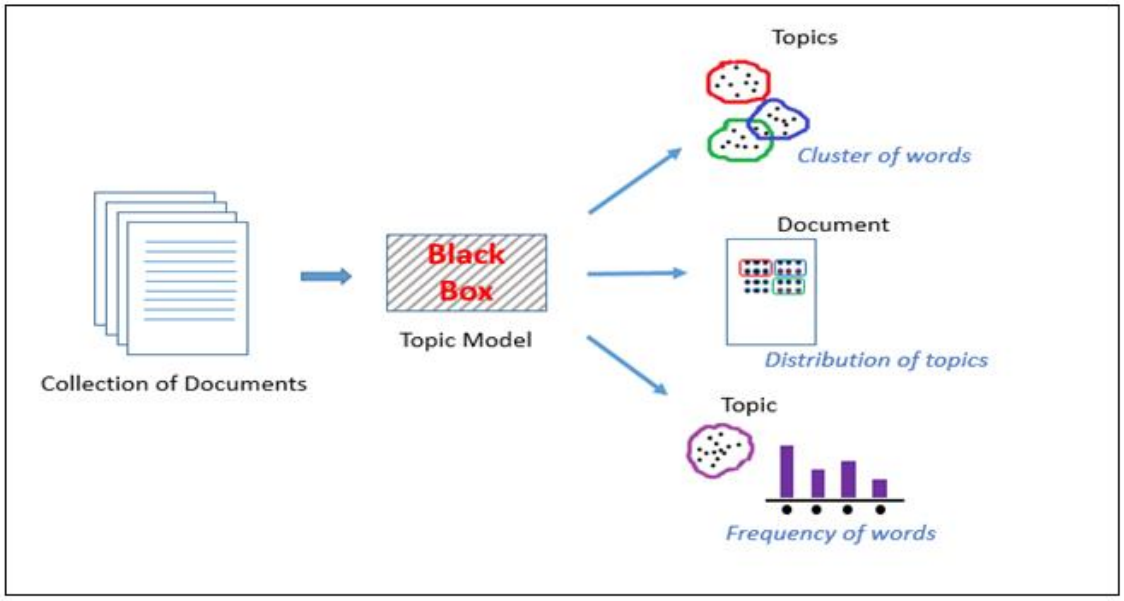

Figure 1: LDA based Topic Analysis

2. Flow Diagram

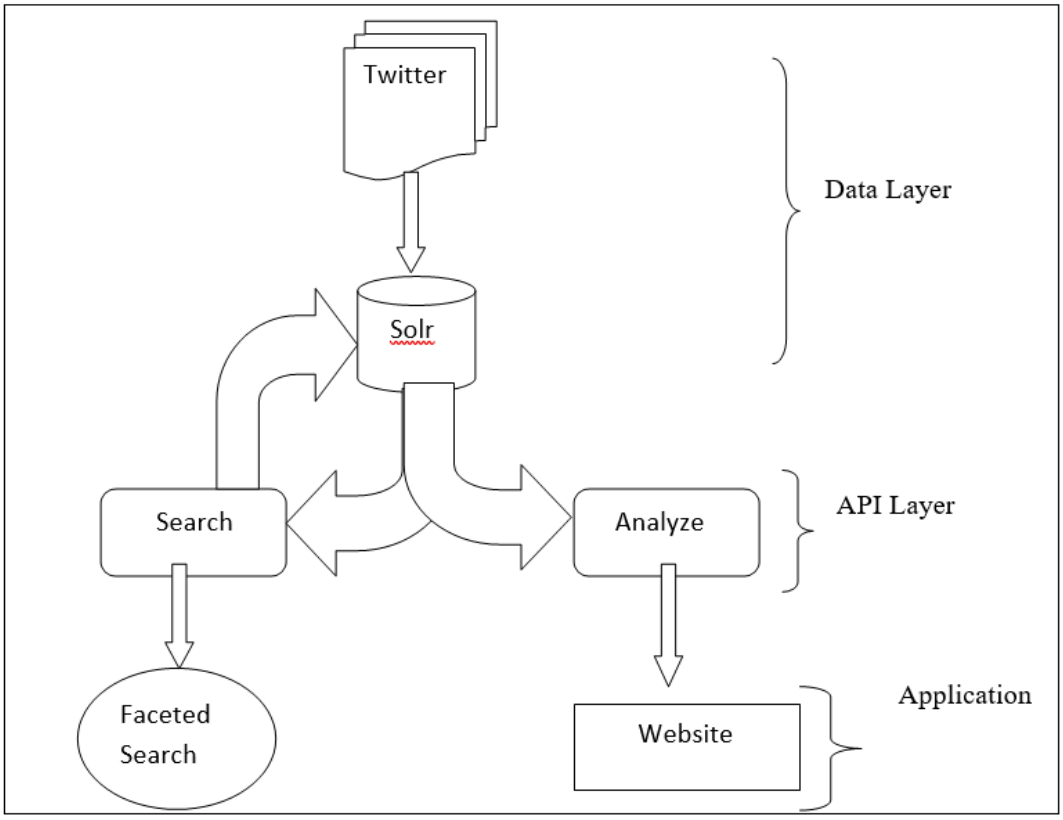

Figure 2: Data Flow Diagram

3. Technology Stack:

3.1 Backend Technologies Used:

- Apache Solr: Apache Solr is an open source search platform built upon a Java library called Lucene. We have used Solr in our project because it can index files and return desired results based on the user query.

- Python Flask: Flask is a lightweight web application framework. It is designed to make getting started quick and easy, with the ability to scale up to complex applications. It began as a simple wrapper around Werkzeug and Jinja and has become one of the most popular Python web application frameworks.

3.2 Front End Technologies Used:

We have used the following languages to design the UI of our website:

- HTML
- JavaScript
- Bootstrap
- Angular JS

3.3 Other APIs Used:

We have performed Sentiment Analysis using TextBlob:

TextBlob: TextBlob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

We implemented Topic analysis using Latent Dirichlet Allocation (LDA) model:

LDA: Topic modelling is a type of statistical modelling for discovering the abstract "topics" that occur in a collection of documents. Latent Dirichlet Allocation (LDA) is an example of topic model and is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modelled as Dirichlet distributions.

4. Results and Reflections:

4.1 Landing page of website

http://sentiment-env.nkq7gv3itk.us-east-2.elasticbeanstalk.com/

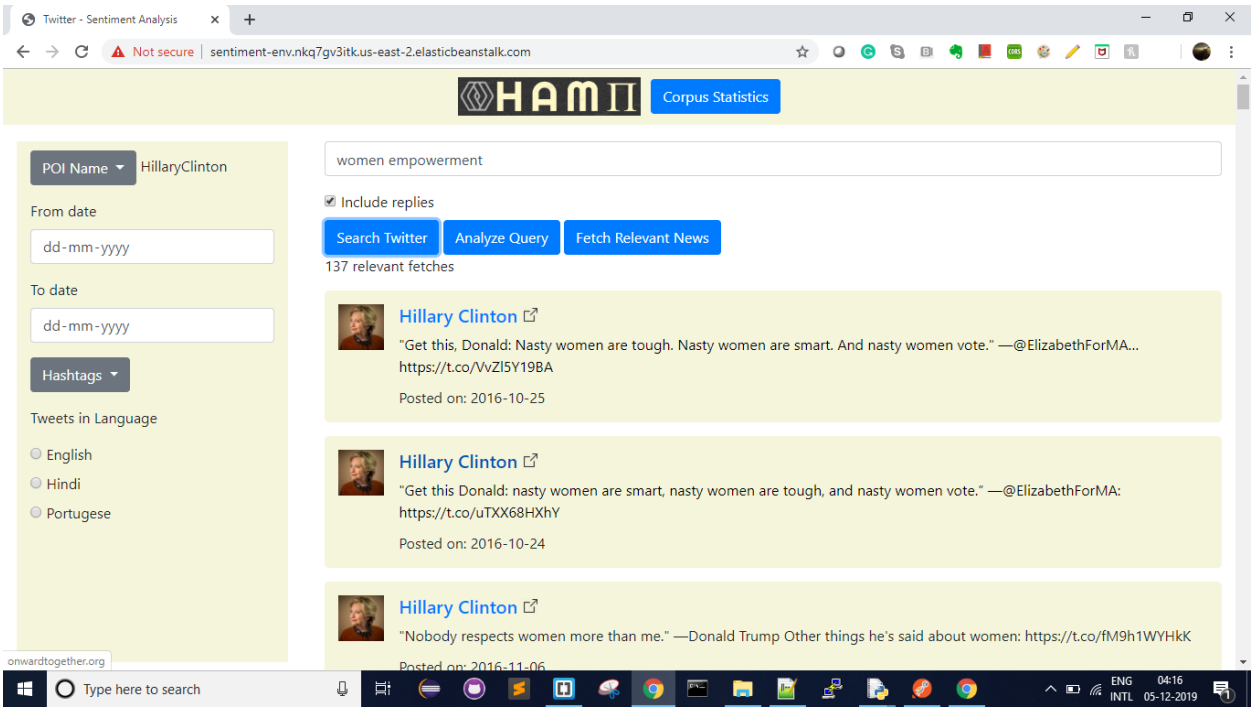

Figure 3: Landing Page

4.2 Graph

4.2.1 Sentiment Analysis and Topic Categorization

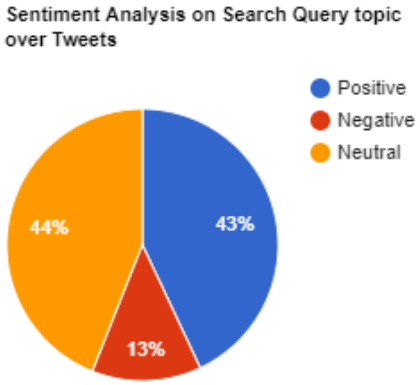

Figure 4: Pie Chart

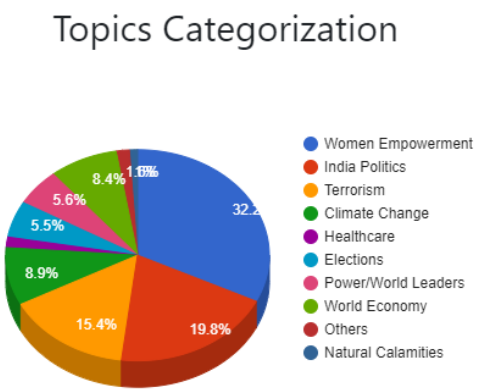

Figure 5: 3D Pie Chart

4.2.2 Top 10 POIs for a particular query search

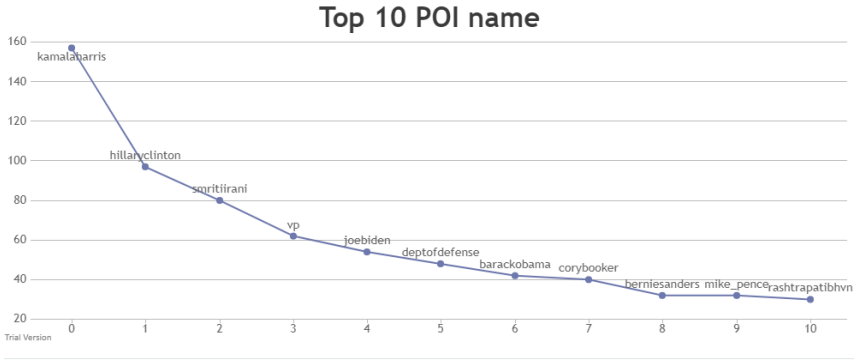

Figure 6: Line Graph

4.2.3 Regional Distribution of tweets on world map

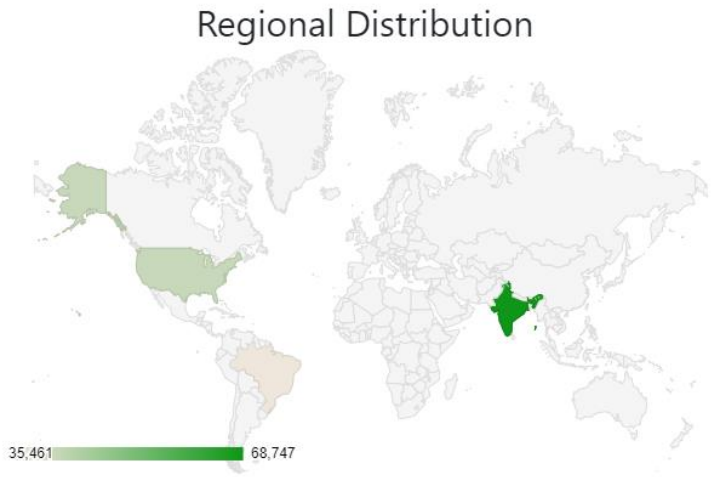

Figure 7: Geographical Distribution of Tweet

5. Group Contribution

Task	Sub-Tasks	Assigned To
Dataset		Astrid
Understanding Solr		Preeti
Designing Workflow		Harshali
AWS		Harshali
Front End:		
	Visualization	Mansi
	Logo	Preeti
	Analytics	Astrid
	Search	Harshali
	Facet Search	Preeti
Back End:		
	Keyword Analysis	Harshali
	Sentiment Analysis	Mansi
	Analytics	Astrid
Fetching News Articles		Harshali
Demo Video		Mansi & Preeti
Report		Mansi & Preeti

6. Conclusion

HAMπ Twitter Analysis Search allows a user to search a particular query on our corpus. We have implemented faceted and query search using Solr. We present sentiment analysis on the query and various line graphs to represent top hashtags, maximum tweet days and POIs with higher tweets on the search query. The analysis on the corpus gives topic categorization, percentage of tweets from the 3 countries with lingual distribution as well as the top 10 hashtags in the corpus. We give relevant news fetches using twitter and POI name as keywords for crawling news on the internet.

7. References

[1] https://lucene.apache.org/solr/6_3_0/index.html

[2] https://textblob.readthedocs.io/en/dev/

[3] https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24

[4] https://www.fullstackpython.com/flask.html

[5] https://docs.aws.amazon.com/ec2/index.html?nc2=h_ql_doc_ec2