CSE535 Project 4: Analyzing the Political Impact of Twitter Data

Abstract

The main objective of our project is to analyze the impact of political rhetoric in traditional and social media. In our project we are extracting data from Twitter and performing analysis of this data based on the person of interest, country, language, etc. We have gathered around 200k tweets which mainly consists of tweets made by impactful people(persons of interest) each having atleast 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 analyze the impact of a tweet by doing sentiment and volume anlaysis.We have displayed most discussed topics from our data collect by doing topic analysis. We have provided keyword search through which we can get tweets and perform analysis of the required data based on the keyword. Search can also be done using a faceted search which is more specific and allows us to get filtered data based on languages,countries,date,etc.

1.Introduction:

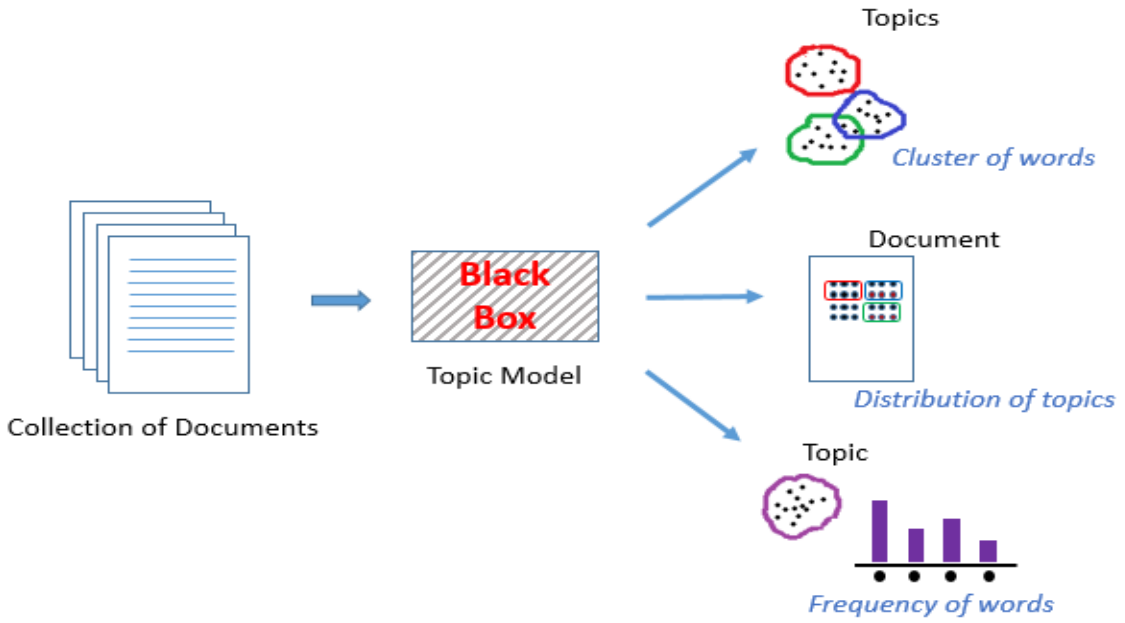
Twitter is an online microblogging tool that disseminates more than 400 million messages per day, including vast amounts of information about almost all industries from entertainment to sports, health to business etc. One of the best things about Twitter—indeed, perhaps its greatest appeal—is in its 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.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.

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 above figure gives us a rough estimate of working of LDA model using which we retrieve main topics from our Twitter data collection.

2.Flow Diagram

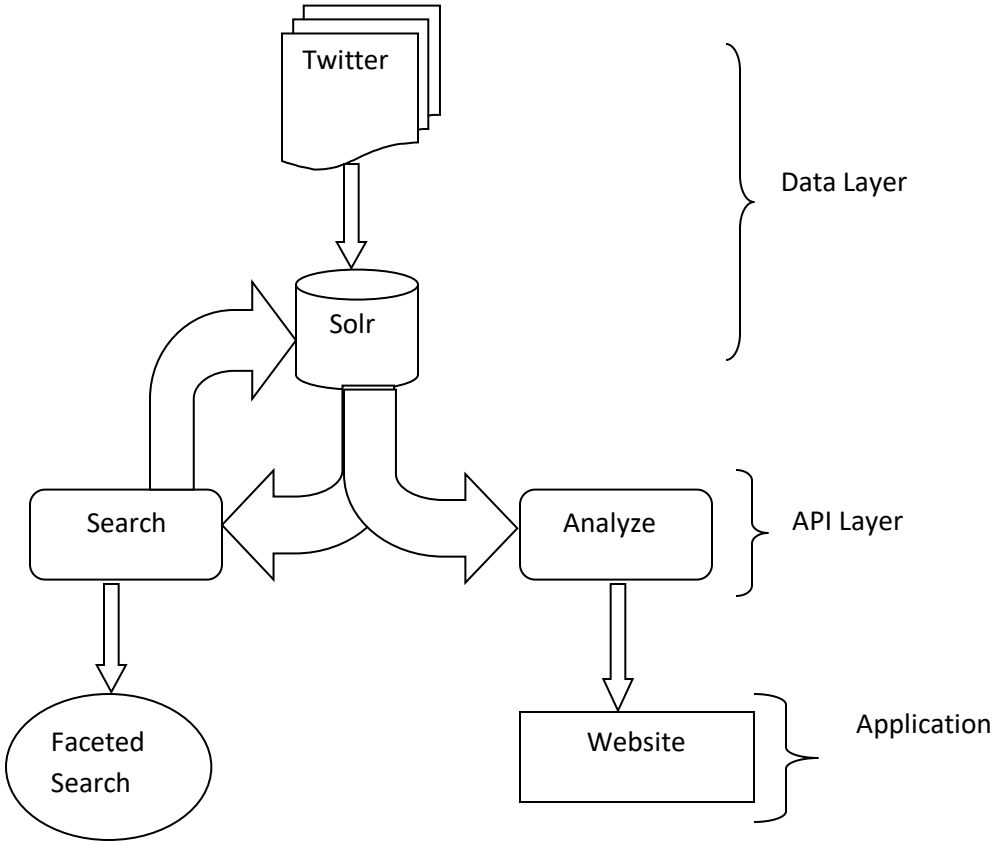


Figure: Data Flow Diagram

4. Technology Stack:

4.1: 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:

4.2: Front End Technologies Used:

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

- HTML
- Javascript
- Bootstrap

- Angular JS

4.3: Other APIs Used:

We have performed Sentiment Analysis:

TextBlob

LDA : nltk, sklearns

5. Results and Reflections:

The Landing page of the website is as follows :