

**Project**

**Twitter Sentimental Analysis of India’s 2019 Election**

**Project Group 2**

Krina Thakkar

Hardik Soni

Yash Lekhwani

Yashashri Shiral

Ritesh Manek

Overview

In India, the General Elections for the Lok Sabha are just around the corner. The news and social media are full of speculation and the numbers and data being generated around this are through the roof.

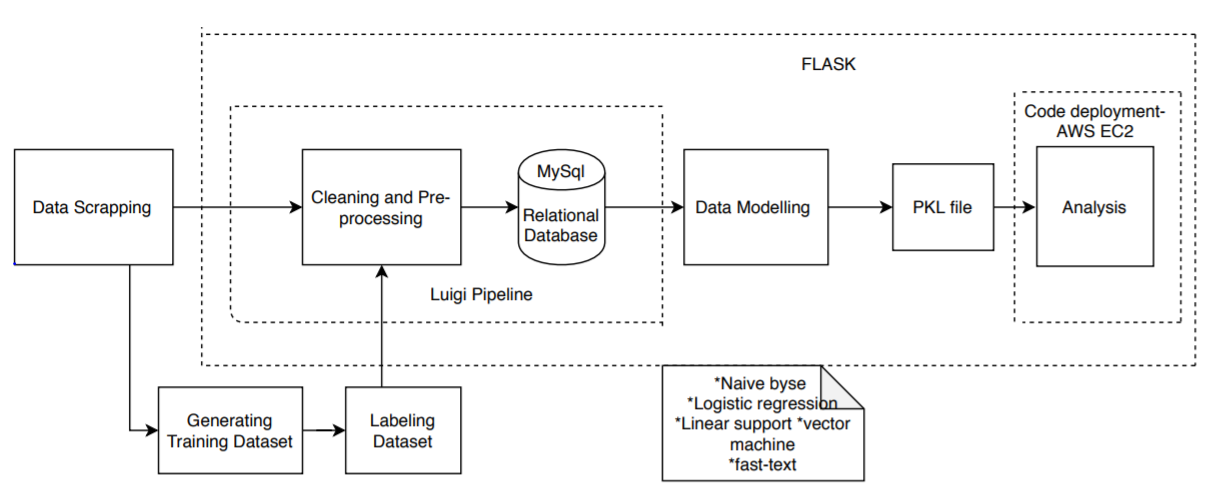
In today’s era of the public being so vocal on the social media, we plan to assess and analyze the political sentiments and understand campaigning and popularity of various political parties by performing a sentiment analysis using tweets, retweet count and location

# Goal

The goal of the Twitter Sentiment Analysis for India 2019 Elections is to:

* Label tweets
* Compare tweet sentiments for both the parties
* Compare parties by understanding the frequency of different tweet sentiments
* Identify commonly used words in tweets
* Understand sentiments by location
* Most popular tweets by retweet count

Pipeline Design



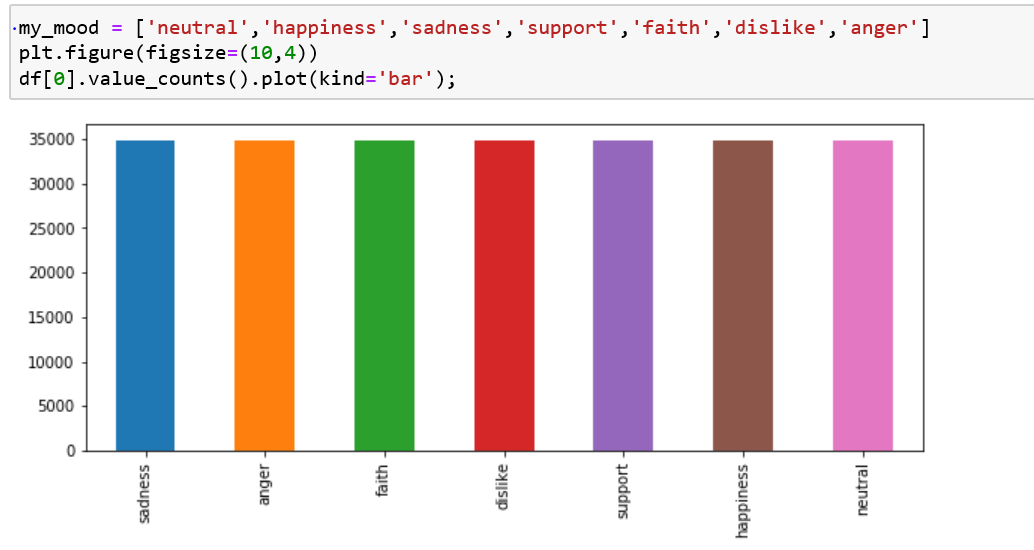
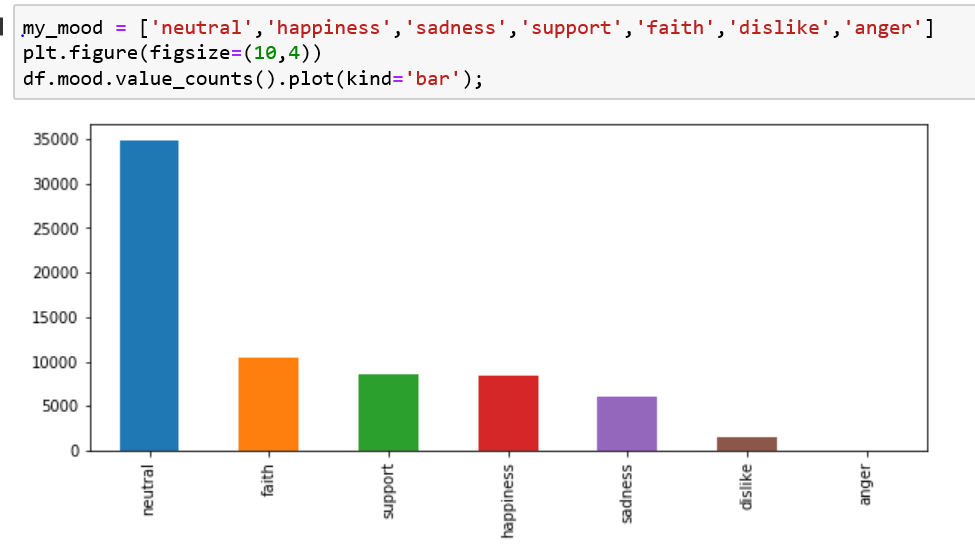
Implementation details

* Data scrapping on twitters weekly data :
  + tweepy- for weekly dataset
  + twitter scrapper - historical Data
  + kaggle - positive and negative dataset for comparison
* In order to handle dependencies and automate the task we used Luigi pipeline
  + Where we get dataset from scrapped data and then perform cleaning and pre-processing operations on the dataset
  + Deployed the code on relational database
* Applying standard machine learning algorithms labelled moods like 'anger', 'dislike', 'faith', 'happiness', 'neutral', 'Sadness', 'support'
* Performed feature engineering on dataset like BJP, Congress and Others
* Stored training and testing dataset in database (RDS)
* Models used: Naïve bayes, Logistic Regression, Linear Support, Fast Text

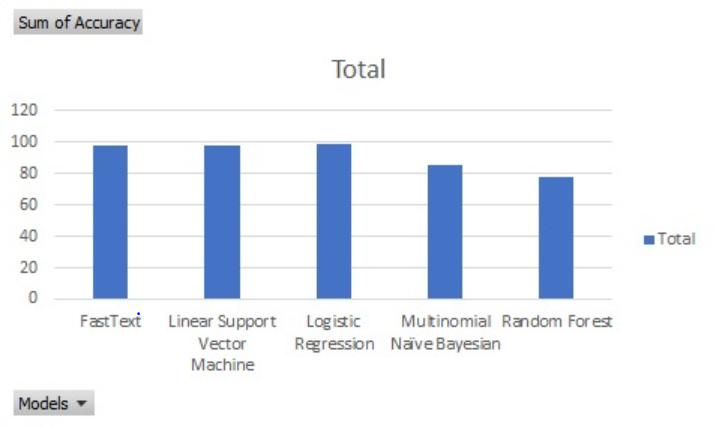
Analysis of Models

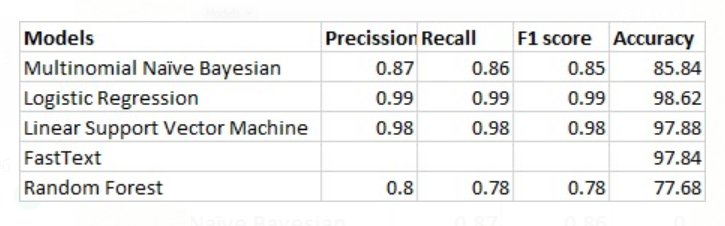
* **Naïve bayes**: Used synthetic minority oversampling technique. performed 5-fold cross validation across all models

After synthetic analysis Before synthetic analysis

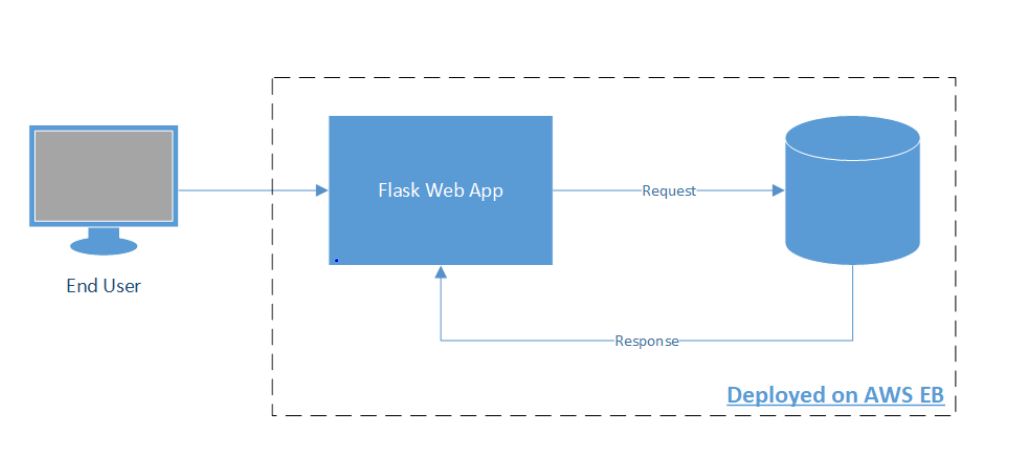
* **Logistic Regression**: Performed 5-fold cross validation. Predict the outcome of dependent variable**Linear Support**: SVM is trained with labelled data
* **FastText**: text based classifiers that helps to analyse character/word at granular level and even provides capability to users to define their own feature-sets from a combination of characters or words





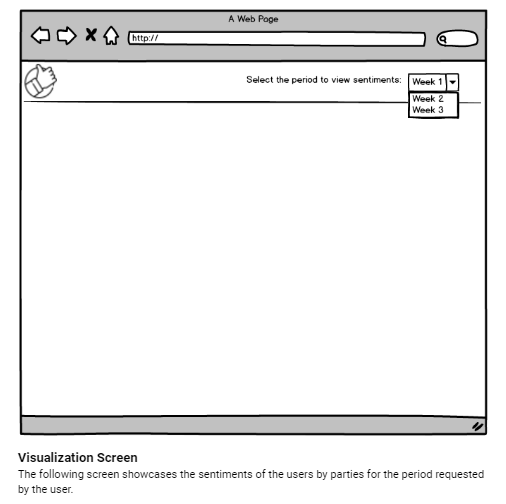
Details on how to run the model

Our final web application will be deployed on AWS.



Steps:

1. Select the time range from the webpage



1. You will see the sentimental analysis of twitter dataset

