Impact Factor (SJIF): 5.301

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 5, Issue 4, April-2018

Political Candidate Popularity Computation Using Sentiment Analysis

Allan Dsouza[1], Divay Mohan[2], Neha Bhamare[3], Diksha Rajput[4], A.M.Jagtap[5]

[1]*Department of Computer Engineering, AISSMS College of Engineering*
[2]*Department of Computer Engineering, AISSMS College of Engineering*
[3]*Department of Computer Engineering, AISSMS College of Engineering*
[4]*Department of Computer Engineering, AISSMS College of Engineering*
[5]*Department of Computer Engineering, AISSMS College of Engineering*

Abstract — ***Candidate popularity is of prime importance especially during or before elections. Traditionally journalists and employees of polling organizations visit the voters to get information about the candidates popularity. Voters when asked about a particular candidate give elaborated answers which contain positive as well as negative sentiments. Due to mixed nature of answers calculating popularity score becomes difficult in traditional approach. Also the calculated popularity score has a very low accuracy .Time,workforce and finance required for this traditional approach is very high whereas results are not satisfactory.***

This project makes popularity computation simple and accurate. Twitter tweets are used to compute the popularity of a leaderor party.Tweets are extracted from twitter app API which will be processed using sentiment analysis and after data from large number of tweets are obtained a popularity score will be computed and displayed.

This will allow journalists and polling organization to take input from bigger voter set and cost for carrying out these tasks will be reduced to a great extent. Users or voters will get the information about popularity in a concise and simple form as it will be displayed in the form of bar graphs and pie charts. Thus easing the task for voters as well as journalists.

I. *Keywords-Sentiment Analysis, Opinion Mining, Natural Language Processing, Text Mining,Popularity Computation.*

I. INTRODUCTION

Opinion Mining or sentiment analysis involves understanding the opinion or sentiment of a person or public as a whole. Understanding the sentiments of a person or public is very important to decide the opinion of people on a particular topic or a decision. Governments can make use of them to understand decision taken by them are being liked or disliked by the public as well as polling organizations can make use of opinion mining in pre-poll election surveys as well as other surveys.

Candidate popularity is of prime importance especially during or before elections. Traditionally journalists and employees of polling organizations visit the voters to get information about the candidates popularity. Voters when asked about a particular candidate give elaborated answers which contain positive as well as negative sentiments. Due to mixed nature of answers calculating popularity score becomes difficult in traditional approach. Also the calculated popularity score has a very low accuracy .Time,workforce and finance required for this traditional approach is very high whereas results are not satisfactory.

In this project we try to simplify the process of popularity computation of political parties. We make use of tweets obtained from tweeter to find the sentiment behind those tweets. We make use of a hybrid model which considers votes from several other algorithms. Finally the results are provided in simple graphical format.

II. Literature Review

Opinion Mining involves a lot of steps ranging from data set collection to sentiment classification. Each and every step can be carried out using different techniques available. As more and more research is being carried out in this field the techniques available are also increasing. Selection of a technique depends upon several parameters such as accuracy, available computational capacity, etc.

There are many stages involved in opinion mining. The first step is collection of data. Data collection is the most important step involved. Without proper data collection opinion mining cannot be started. The next step involves cleaning and stemming of words present in the data set. Cleaning and stemming are also important to improve the overall accuracy of the opinion mining task as it involves removal of unnecessary words and conversion of words to more relevant words.

The next task is application of the opinion mining algorithm. Also due to research carried out on extensive scale we have many different techniques available at our disposal. These techniques involve various machine learning techniques like Naïve Bayes classification,support vector machine classification and a wide range of corpus based techniques such as TF (Term Frequency), IDF (Inverse Document Frequency), etc.

The last step involves checking the accuracy of the prescribed algorithm for the given data set. The accuracy may vary across data sets and types of data involved and checking the accuracy is very important to decide the current algorithm is useful for the current data set or not.

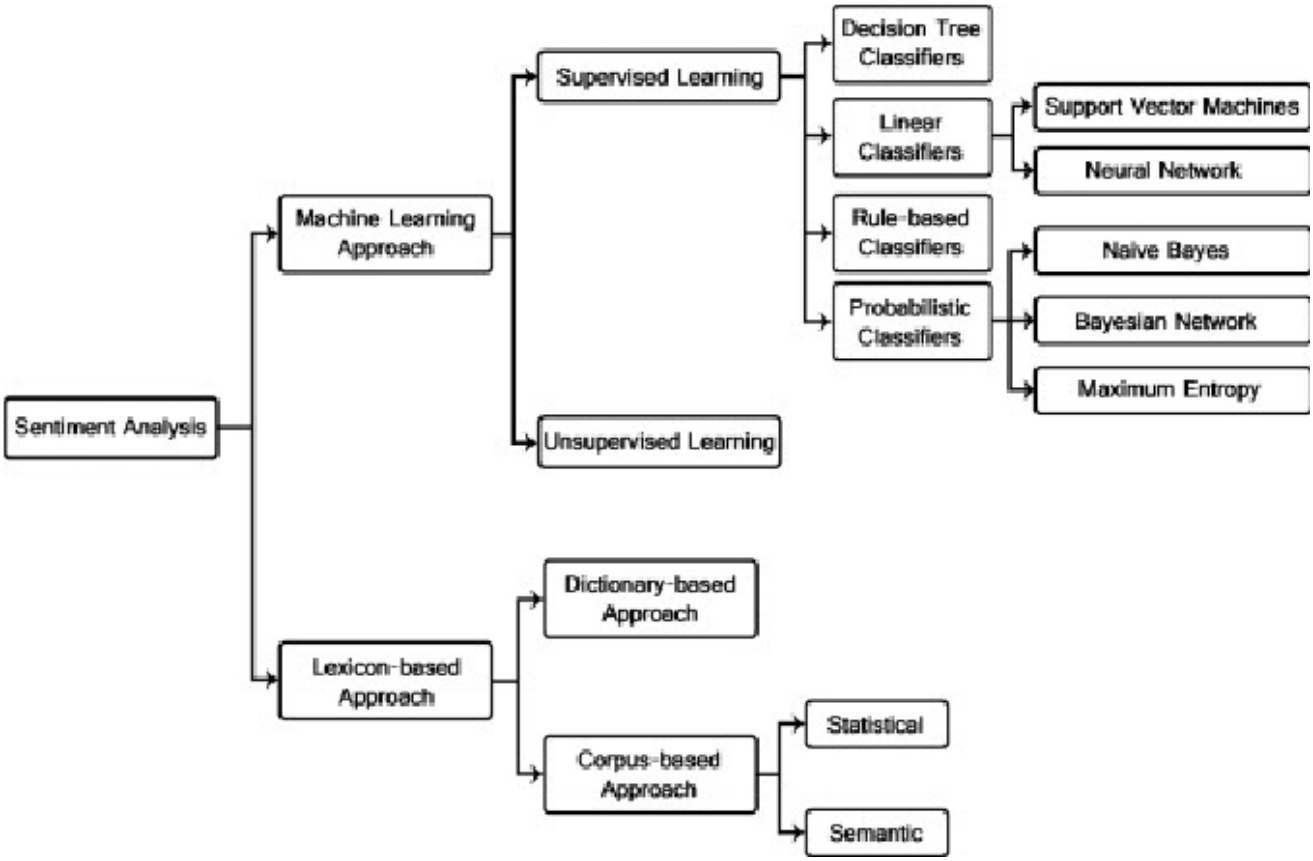


Figure 1. Commonly used techniques for sentiment analysis

III. Methodology

Step 1: Data Collection

In this project we make use of twitter tweets for sentiment analysis .In order to collect these tweets we are using Twitter App API. We login through the Twitter App API using our own username and password. The App API generates consumer keys. These Keys Are required in our programs which collects the tweets from twitter. We make use of a program which accepts a keyword and collects relevant tweets. This program takes the keys obtained from twitter app API to collect the tweets related to given keyword. These Tweets are then stored in a json file. We then convert this json file filled with Tweets into CSV file for further processing. .But the problem with the collected data is that it has lot of impurities like codes for emojis, links etc. So before applying the model data preparation is necessary.

Step 2: Data Preparation

Data preparation involves the following:

a)Data Cleaning by Regular Expression like emojis.
b)Tokenization of data.
c)Prepare Feature set.

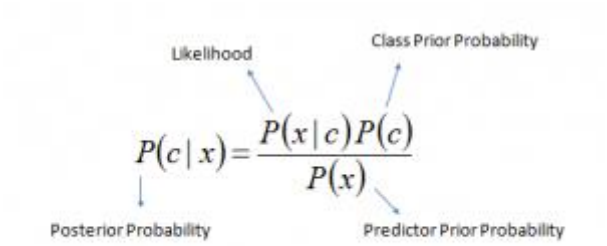
Step 3: Train Model

We used training data of 10000 data point(tweets) which are already binary classified. Out of which 9400 sentiments used for training model.

I. **Naïve Bayes:**

Naive Bayes is a very popular model in the field of text classification sentiment analysis. It is a classification technique based on Bayes' Theorem with an assumption of independence among predictor. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look the equation below:



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

II. Gaussian Naive Bayes

Gaussian Naïve Bayes implements the Gaussian Naive Bayes algorithm for classification. The likelihood of the features is assumed to be Gaussian:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

The parameters σ_y and μ_y are estimated using maximum likelihood.

III. Multinomial Naïve Bayes:

Multinomial Naïve Bayes implements the naive Bayes algorithm for multinomially distributed data, and is one of the two classic naive Bayes variants used in text classification (where the data are typically represented as word vector counts, although tf-idf vectors are also known to work well in practice). The distribution is parametrized by vectors $\theta_y = (\theta_{y1}, \ldots, \theta_{yn})$ for each class y, where n is the number of features (in text classification, the size of the vocabulary) and θ_{yi} is the probability $P(x_i \mid y)$ of feature i appearing in a sample belonging to class y.

The parameters θ_y is estimated by a smoothed version of maximum likelihood, i.e. relative frequency counting:

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

where $N_{yi} = \sum_{x \in T} x_i$ is the number of times feature i appears in a sample of class y in the training set T, and $N_y = \sum_{i=1}^{|T|} N_{yi}$ is the total count of all features for class y. smoothing priors $\alpha \geq 0$ accounts for features not present in the learning samples and prevents zero probabilities in further computations. Setting $\alpha = 1$ is called Laplace smoothing, while $\alpha < 1$ is called Lidstone smoothing.

IV. Logistic Regression

Although Naïve Bayes and Logistic regression are linear classifiers, Naïve Bayes is a generative classifier that tries to predict the likelihood term under the assumption of conditionally independent between features, however,this assumption less happened in the real word problems,but by contrast, Logistic regression is a discriminative classifier that uses a Logistic function to get the likelihood directly. In addition, Logistic regression covers the case of a binary dependent variable. As our dataset are going to be classified in a positive and negative categories, it is expected to have better performance than Naïve Bayes.

V. Linear SVC

As the Logistic regressors do not optimize the number of mislabeled data, we need another classifier to minimize the misclassification error rather than solely relies on the likelihood. Therefore,the support vector machines model with a classifier of the form shown in Eq. (1) is chosen. It can minimize the misclassification error for prediction as written in Eq.(2):

$$y = \{1 \text{ if } x_i \theta \alpha > 0 ; \ 1 \text{ otherwise}\} \quad (1)$$

$$\text{a rgmin } \theta \sum \delta (y_i (x_i \theta \alpha) \leq 0) \quad (2)$$

Support Vector Machine constructs a set of hyperplanes in a high dimensional space for classifications. In order to achieve a good classification, the distance from hyperplanes to the nearest training points should be maximized. As in general, the larger the margin,the lower the generalization error of the classifier. We use C type Support Vector Classification with penalty parameter C of the error term to solve Eq.(3):

$$\min w, b, \varsigma \, 21 \, w^t w + C \sum \varsigma_i \quad (3)$$

this equation is subjected to $y_i (w^t \varphi(x_i) + b) \geq 1 \varsigma I$,where $\varsigma_i \geq 0$, $i = 1, \ldots, n$ and C as a regularization parameter,which is defined as C = Number of Samples/alpha.

VI. Hybrid Model

We are making use of hybrid model as our main model. Hybrid model makes use of all the above models. It takes votes from all the above models and then provides final results.

IV. **Results**

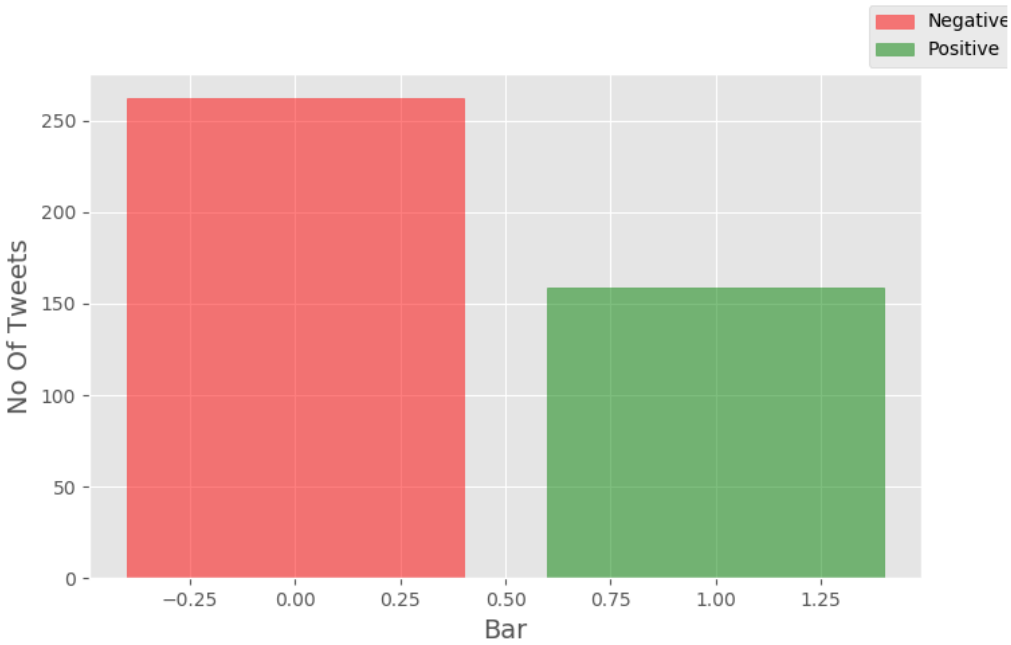
Testing Models:

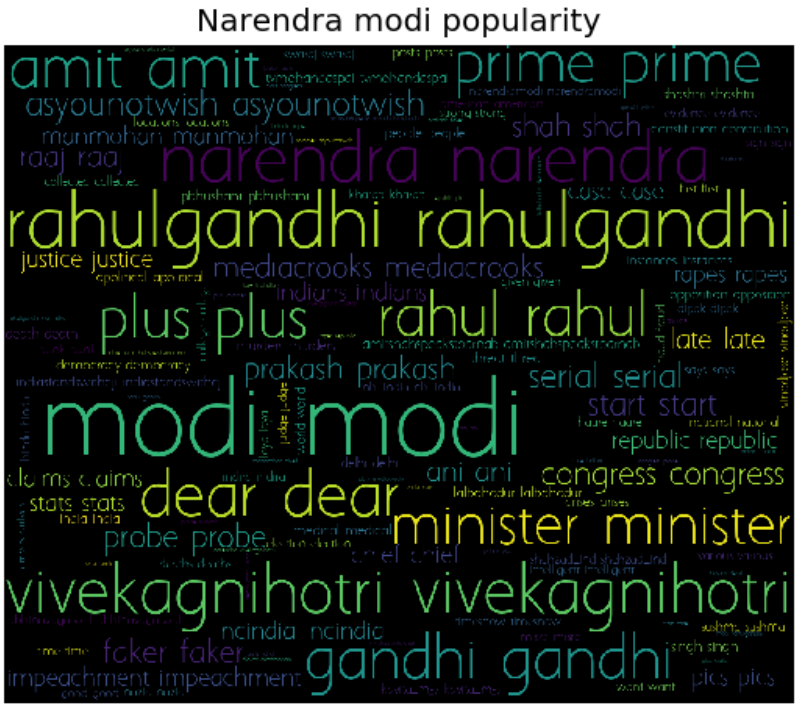
After Training the models , each model was tested on 600 tweets the result as follows:

Model	Accuracy (%)
Naïve Bayes	72.78
Multinomial Naïve Bayes	71.38
Bernoulli Naïve Bayes	71.53
Logistic Regression	73.19
Support Vector Machine(SVC)	70.33
Hybrid Model	75.15

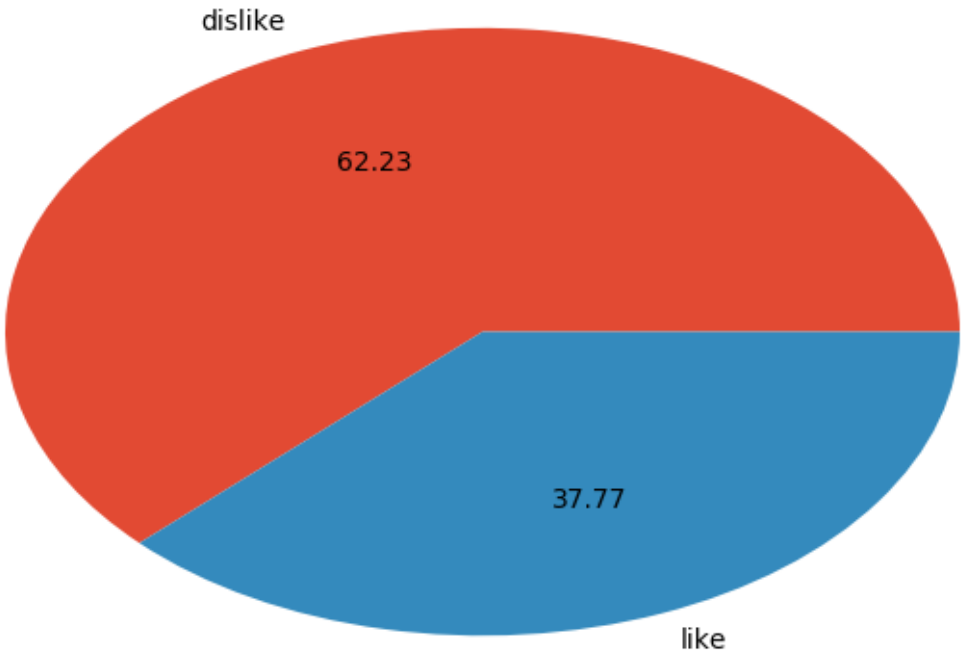
Data Classification results

For Sample We Collect The data with keyword “BJP” (Bhartiya Janata Party) And We get result as Follow:





Top 5 words:



Like and Dislike

	terms	frec
0	modi	5338
1	@rahulgandhi	4987
2	narendra	4115
3	@vivekagnihotri	4063
4	dear	3815

Term Frequency

V. Conclusion

Opinion mining or sentiment analysis is a widely growing field of Natural Language Processing. It is truly a multidisciplinary field due to applications in different domains. Sentiment analysis also possesses challenges which ranges widely. Overcoming these challenges is very important to ensure high accuracy and proper usability of the results obtained through it. Twitter is a increasing popular and widely accepted platform .Twitter tweets are also available freely for usage .This makes use twitter tweet analysis as the best way for finding the popularity of candidates. Increasing amount of research in this field of Sentiment analysis has introduced new and unique algorithms and techniques for sentiment analysis, these techniques ensure higher accuracy and simplicity. Availability of these different techniques and their successful usage by different researchers provide us with wide variety of techniques at our disposal. A hybrid model which uses all these techniques can be considered best due to higher accuracy and more reliability.

REFERENCES

[1] Anshuman, Shivani Rao, Misha Kakkar," Rating Approach Based on Sentiment Analysis",2017, IEEE.

[2]. Harshali P Patil, Dr Mohammad Atique," Sentiment Analysis for Social Media: A Survey",2015, IEEE.

[3] Neethu M S, Rajashree R," Sentiment Analysis in Twitter Using Machine Learning Techniques",4 July 2013,4th ICCCNT.

[4] MS K Mouthami, Ms K Nirmala Devi, Dr. V Murli Bhaskaran," Sentiment Analysis and Classification Based on Textual reviews

[5] Zhu Nanli, Zou Ping, LeeWeiguo, Cheng Meng," Sentiment Analysis: A Literature Review",2012, IEEE.

[6] Khaled Ahmed, Neamat El Tazi, Ahmed Hany Hossny," Sentiment Analysis Over Social Network: An Overview,2015, IEEE.

[7] Monisha Kanakaraj, Ram Mohana, Reddy Guddeti," NLP Based Sentimental Analysis on Twitter Data Using Ensemble Classifiers,2015, IEEE.

[8] Santanu Mandal, Sumit Gupta," A Lexicon-Based Text Classification Model To Analyze And Predict Sentiments From Online Reviews".

[9] Bo Pang, Lillian Lee, Shivkumar Vaithyanathan," Thumbs up? Sentiment Classification Using Machine Learning Techniques".

[10] Shivprasad T K, Jyothi Shetty, "Sentiment Analysis of Product Reviews: A Review",2017, IEEE.