Twitter Sentiment Analysis for Classifying Hate Tweets and Normal Tweets Using Logistic Regression and Naive Bayes Algorithm.

[1]Swati Powar, [2]Unmesh Kadam, [3]Tanmay Salvi
[1]Assistant Professor, [2]Student BEIT, [3]Student BEIT
[1] Department of Information Technology,
[1]Finolex academy of management and technology Ratnagiri, India

Abstract : Twitter keeps record of all the tweets posted by millions of users all over the globe. This data stores a number of attributes of the tweet as well as the individual registered user. This data is stored together in Twitter Datacentre. This data is mined for detecting frequent patterns as well as anomalies and is currently used to rate a particular product or movie. This data can also be used for classifying any particular tweet as Hate tweet and Normal tweet using various Machine Learning techniques. From the literature survey, it has been found that Logistic Regression provides less accuracy in solving the classification problems. Hence, we can use Naive Bayes method for accuracy improvement.

IndexTerms **- Classifying Tweets, Logistic Regression, Naive Bayes, Machine Learning.**

I. INTRODUCTION

"We live LESS in real and MORE in virtual, this statement is turning out to be a reality with the rapid development of the Internet and Social media platforms and the ultimate increase in the number of social media users, the issue of fake news becoming viral and thus disturbing the law and order situation of the entire nation is becoming more serious day by day. Organizations like Facebook, Twitter are trying to identify and censor problematic posts while weighing the right to freedom of speech. Due to the cheap availability of computing and storage, it has become possible to use sophisticated machine learning algorithms for this purpose. In this paper, we are providing solution to classify a particular tweet as hate tweet or normal tweet based on the historical data of combination of various tweets volume. According to the characteristics of the data, we can use the method of logistic regression and Naïve Bayes to classify the tweets

II. LITERATURE SURVEY

The utility of linguistic features for detecting the sentiment of Twitter messages. We use three different datasets - Hash tagged, Emoticon, in the experiments[1]. A tweet is offensive if ,it uses a sexist or racial slur,attacks a minority,seeks to silence a minority.criticizes a minority (without a well founded argument),promotes but does not directly use, hate speech or violent crime[2]. By iterating through the training set, Naive Bayes classifier finds out the number of occurrences of each bigram word and checks if the test sentence has the same feature words as the training data. After the preprocessing of training set is completed, the bigram feature vectors are extracted from every tweet[3].

III. DATA

For training the classification models, we are using a labelled dataset of 31,962 tweets. The dataset is in the form of csv file. Each record in the dataset contains following attributes

1] tweet id
2] tweet label
3] tweet

The used Dataset has 29720 tweets labelled as ' 0' and 2242 tweets labelled as ' 1' .The memory usage of the Dataset is 749.2 KB.

IV. METHODOLOGY

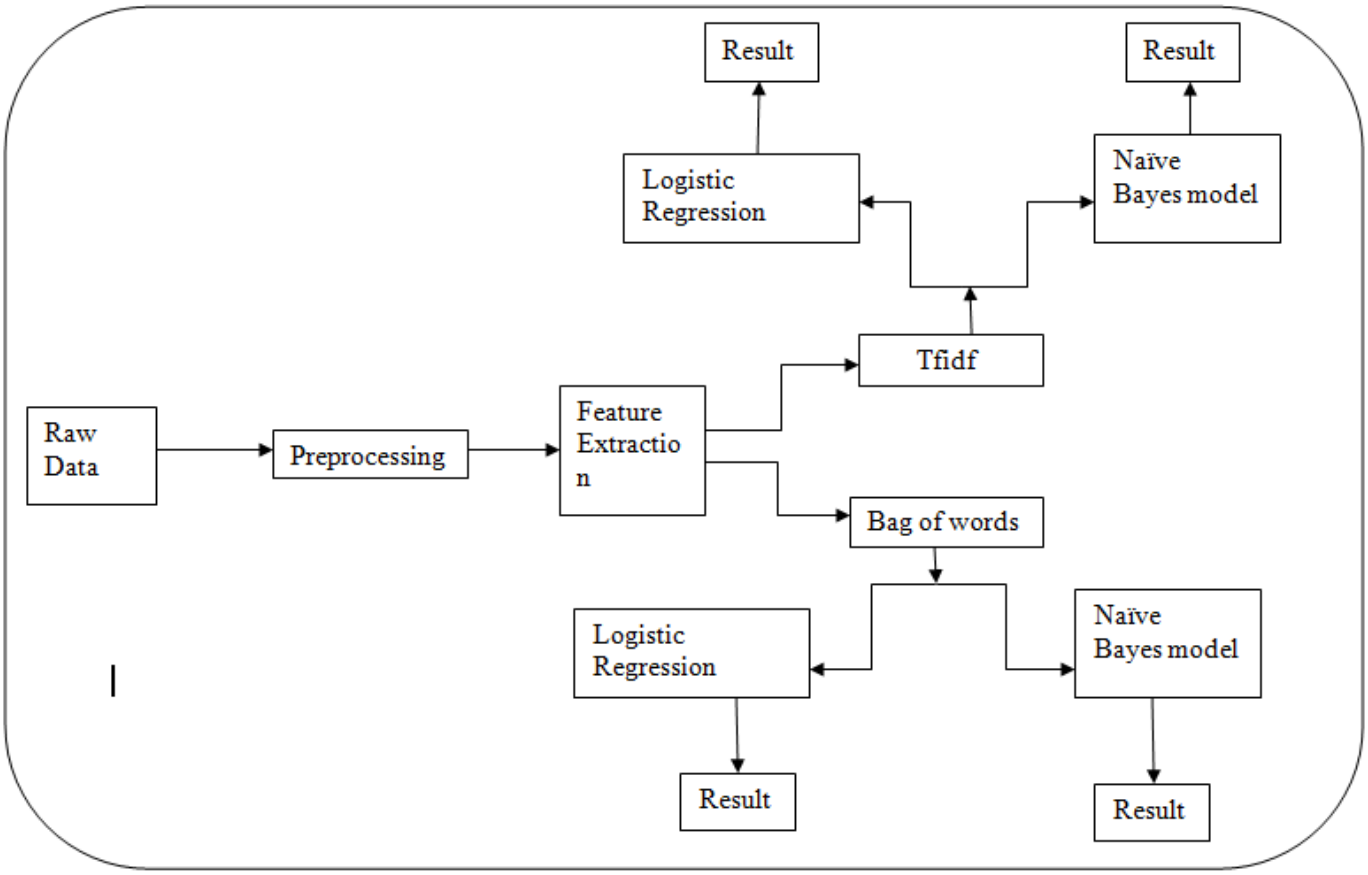

fig1.Architecture diagram**3.1Population and Sample**

We propose the above methodology for solving the problem. First we perform data pre-processing for noise removal which includes discarding language stop words URLs or links, mentions, hash tags, special characters and word stemming. Once we are done with pre-processing, then we perform feature extraction using bag of words (BOW) and Term frequency- inverse definition frequency (TFIDF). Then we use logistic regression classifier and naive bayes classification algorithms on both of the extracted features i.e. BOW and TFIDF to classify the tweets.

4.1 Logistic Regression

Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function. The recommended sample size for each category of independent variable is at least 10 observations per estimated parameter. Logistic regression uses large sample size which decreases the chances of errors

$$Y = b0 + b1X1 + b2X2 + ...bkXk-E \quad (1)$$

Y= Dependent variables b= coefficient of variable X X= independent variable E= error term

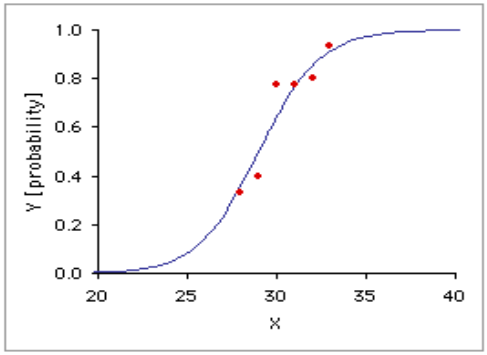

Fig. 2 Logistic Regressions.

4.2 Naïve Bayes

It is a classification technique based on Bayes Theorem. Naïve Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. It is easy to predict class of test data set.

$$P(c \mid x) = P(x \mid c)\, P(c) / P(x) \quad (2)$$

P(c | x) = posterior probability of the target class
P(c) = prior probability of class
P (x | c) = likelihood which is the probability of predicator class
P(x) = prior probability of predicator

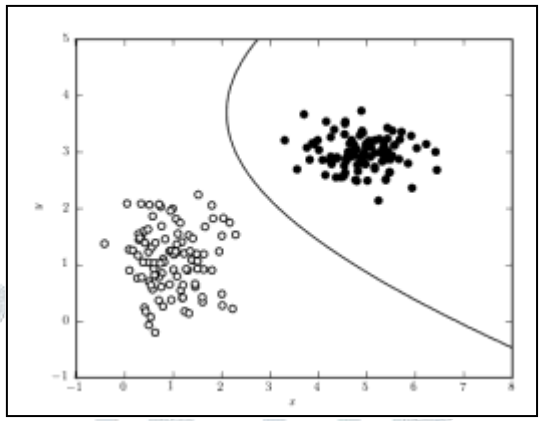

Fig. 3 Naive Bayes.

V. CONCLUSION

Hence, we propose in this paper a system for classifying the tweet as hate twitter normal tweet based on the twitter dataset. Using this system the accuracy of prediction for logistic regressions model and naive bayes model can be determined

REFERENCES

[1] Ms. Swati Powar, Dr. Subhash Shinde " Named Entity Recognition and Tweet Sentiment Derived From Tweet Segmentation using Hadoop." ,IEEE Mar.2017

[2] Mitali Desai, Mayuri Mehta " Techniques for Sentiment Analysis of Twitter Data: A Comprehensive Survey , International Conference on Computing, Communication and Automation (ICCCA2016)

[3] Geetika Gautam, Divakar Yadav " Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis" , IEEE Dec.2014.