Machine Learning Techniques for Sentiment Analysis of movie reviews

B.Bhargav Srikar[1], Vaibhav Yalla[2], Dr.P.V.Lakshmi[1]

[1]Department of IT, GITAM Institute of Technology, GITAM, Visakhapatnam
[2]Department of CSE, GITAM Institute of Technology, GITAM, Visakhapatnam
E-mail: sreekar619@gmail.com, pvl.7097@gmail.com@gmail.com, yallavaibhav@gmail.com

Abstract

In this smart world, every person has a chance to express their opinion in a digital way. Earlier, if a business executive wants to get feedback from the customer, he would have to find a means to know it from the customer directly. But with the explosion of social media, anyone today can express their views about anything in an unconstrained and freeway. In this paper, Naive Bayes and KNN models are developed to find the sentiments of people writing reviews with a goal of understanding the opinions of the public on a movies. Also, the emotions of people associated with different comments and their corresponding percentages have been calculated. The performance and accuracy of Naïve Bayes model is significantly good.

.

Index Terms- *Sentiment analysis, machine learning, supervised machine learning*

1. Introduction

Sentiment Analysis is one of the most learned research areas for prediction and classification. Automated Sentiment Analysis of text is used in fields where products and services are reviewed by customers. Beginning from being an archive level classification [1], it has been taken care of at the sentence level [2,3] and then at the expression level [4,5]. Analysis helps concerned organizations to find opinions of people about movies from their reviews. The reviews given by the customers and critics about the movies and their services are in the form of text analysis. This type of Sentiment Analysis helps clients to improve their business based on the reviews given by the customers. Two Models are built on the extracted data for analysis of Positive and Negative sentiment.

2. MODEL BUILDING AND EVALUATION

2.1 The proposed work provides a comparison on sentiment analysis of IMDB movie data using Machine learning methods Naïve Bayes and KNN. Python code is used to model the algorithm.

2.2 Dataset Description

We have taken IMDB dataset consisting of 50000 reviews from IMDB Database. Each of these movie reviews is classified as either “positive” or” negative”. The data is divided into a train and validation set, each with 50% positive and 50% negative reviews. From all movie reviews only the top 20000 most frequently occurring words are used. Reviews have been transformed into sequence of integers. We have developed Naïve Bayes model and KNN model. 70% data is trained and rest 30% data is used for testing. The confusion matrix is obtained for summarizing the performance of both the models. We have calculated precision, recall, f1-score and accuracy for both positive and negative reviews. Also we have taken amazon review data set from kaggle to calculate the accuracy of test data using our models.

2.3 Models

Naive Bayes Model

Naive Bayes technique is based on Bayes’ theorem with an assumption of independence when the dimensionality of the inputs is high. Naive Bayes can often outperform the sophisticated classification methods. Mathematically, Bayes' theorem is stated as

$$P(A|B) = P(A).P(B|A) / P(B)$$

where A and B are events and P(A|B) is probability of A given B.

Popular Naive Bayes classifiers are Multinomial Naive Bayes and Bernoulli Naive Bayes. Multinomial Naive Bayes: In Multinomial Naïve Bayes model, feature vectors represent the frequencies with which certain events have been generated by a multinomial distribution. This model is typically used for document classification.

Bernoulli Naive Bayes: In this model, features are independent booleans describing inputs. This model is popular for document classification tasks, where binary term occurrence features are used rather than term frequencies and $P(B) \neq 0$.

The generated Naive Bayes model gave test accuracy of 85.84%.

KNN Model

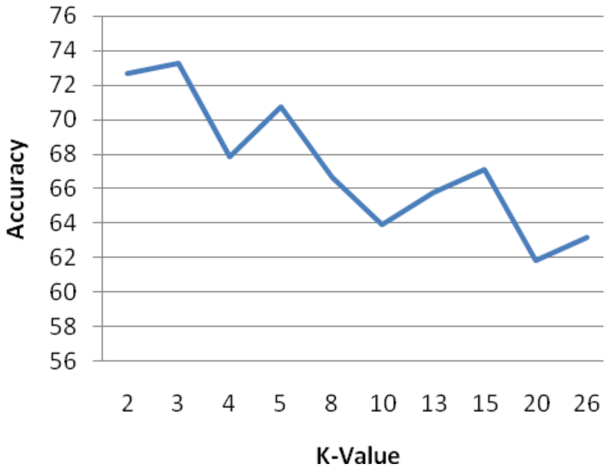
KNN model doesn't have a specialized training phase, it is also known as lazy learning algorithm. Rather, it uses all of the data for training while classifying a new data point or instance. KNN is a non-parametric learning algorithm. This is an extremely useful feature since most of the real world data doesn't really follow any theoretical assumption. We used Euclidean distance method to calculate the distance of a new data point to all other training data points. The generated model gave test accuracy of 75.24% for k equal to 3.

2.4 Evaluation

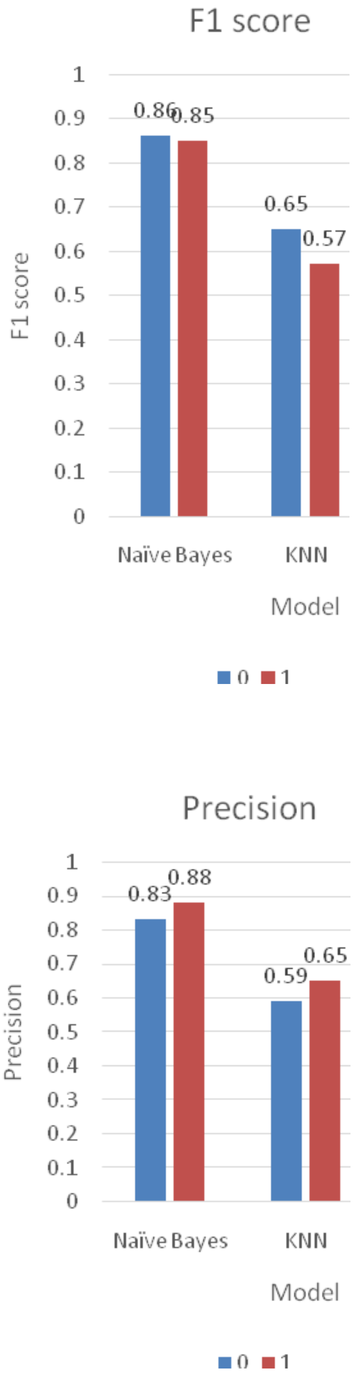
For evaluation, we used the popular metrics such as precision, recall, F1-score and accuracy.
Precision = truepositives /(truepositives +
Falsepositives)
Recall = truepositives/(truepositives+
Falsenegetives)
F1 Score = 2 × (Precision × Recall) / (Precision
+ Recall)
Accuracy=truepositive+truenegetive/
(truepositive + false negative
+false positive + truenegetive)
The metrics give an option to compare the models among the methods based on different parameter changes.

3. RESULT AND ANALYSIS

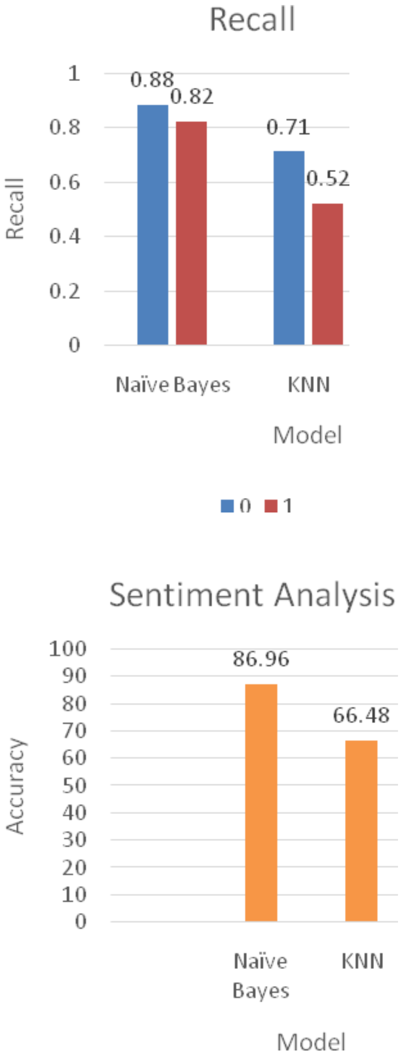
The experimental results is obtained for identifying the sentiment of IMDB using Naïve Bayes and KNN model. In order to find the suitable K value for KNN model, the experiment with KNN was run for cross-validation for each $K = \{1, 2, . . . , 26\}$ with 30% data considered in the test case. Maximum accuracy of 73.24% is obtained by KNN with value k=3. The F1 score, Precision, Recall and accuracy of IMDB data set using Naïve Bayes and KNN models are shown below, where 0 and 1 are negative and positive sentiments. It is observed that Naïve Bayes model, compared to KNN model has shown an improved score. Test accuracy of Naïve Bayes model comes out to be 85.84%.



KNN model accuracy: For k=3 the accuracy is maximum.



F1 score and Precision of Naïve Bayes and KNN model



Recall and Accuracy of Naïve Bayes and KNN model.

4. CONCLUSION

In this paper, the Goal is to calculate the polarity of sentences that can be extracted from the text of reviews. The model is built from movie reviews and tries to find out the sentiment matches for these movies. A comparison on sentiment identification of IMDB using NB and KNN is shown. Naive Bayes algorithm and KNN model has been used to find the sentiments of people writing reviews with a goal of understanding the opinions of the public on movies. The performance and accuracy of Naïve Bayes model is significantly good. This will help Film makers and Amazon sellers to check the status of their movies and products.

References

[1] Bo Pang and Lillian Lee," A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts "ACL '04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics article no 271.(2004).

[2] Minqing Hu and Bing Liu "Mining and Summarizing Customer Reviews" KDD '04 Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining Pages 168-177, (2004)

[3] Soo-Min Kim, Eduard Hovy, "Determining the Sentiment of Opinions" COLING '04 Proceedings of the 20th international conference on Computational Linguistics, Article No. 1367 (2004).

[4] Marenda A. Wilson, Stacie Meaux, Roy Parker, and Ambro van Hoof "Genetic Interactions between [PSI^{+}*] and nonstop mRNA decay affect phenotypic variation" PNAS 2005 10244 - 10249*

[5] Apoorv Agarwal, Boyi Xie,Ilia Vovsha ,Owen Rambow ,Rebecca Passonneau "Sentiment Analysis of Twitter Data" Proceedings of the Workshop on Language in Social Media(LSM 2011), pages 30– 38, Portland, Oregon, 23 June 2011. c 2011 Association for Computational Linguistics.