# "IMPROVING ACCURANCY OF TEXT CLASSIFICATION USING BAYESIAN ALGORITHMNON 20 NEWSGROUP DATASET"

**[1]PRASHANT G. GHULAXE**
**ME Scholar, Department of Computer Science & Engineering, HVPM College of Engineering and Technology, Amravati, India**
**pghulaxe@gmail.com**

**[2]DR. ANJALI B. RAUT**
**Head, Department of Computer Science & Engineering, HVPM College of Engineering and Technology, Amravati, India**
**anjali_dahake@rediffmail.com**

**ABSTRACT:** *Text classification deals with allocating a text document to a predetermined class. Generally, this involves learning about a class from representations of documents belonging to that class. Documents to be classified are commonly represented by a bag-of words feature vector. The bag-of-words model cannot handle two language phenomena: synonymy and poly se my; besides, dimensions of feature vectors are orthogonal. In order to effectively address those problems, some researchers adopt a bag of-concepts representation of documents—understanding concept as "unit of meaning". In this paper we proposed the application that can effectively classify text files into appropriate folder depending upon the theme of the file, using the training data to model the classifier. This application automates the text classification process otherwise would take long time doing manually the same task. Text file are appropriately classified using this application.*

***Keywords****:* Text classification, bag-of-words,

## 1. INTRODUCTION

Automated text classification is a particularly challenging task in modern data analysis, both from an empirical and from a theoretical perspective. This problem is of central interest in many internet applications, and consequently it has received attention from researchers in such diverse areas as information retrieval, machine learning, and the algorithms. It has been widely observed that feature selection can be a powerful tool for simplifying or speeding up computations, and when employed appropriately it can lead to little loss in classification quality. A major part of this information is stored in text databases and repositories which contain enormous set of documents belonging to various sources like research papers, email messages, news articles, entertainment, art, literature, books, web pages and digital libraries. In most of organizations nowadays, information is kept in the form of text databases like in health, government, education, business and various other areas. The text stored in the databases is in inherently unstructured or semi structured format. The growing enormity and complexity of this unstructured text data has made it a vital need to manage it effectively. This can be done by classifying the documents to simplify the storage, retrieval and presentation of text data. Manual data analysis has become very tedious due to large dimensionality of text data. In such a case, text mining is the only hope to manage this enormous collection of data.

## 2. RELATED WORK

Amna Rahman et. al. [1] proposed a classifier combination that uses a Multinomial Naive Bayesian (MNB) classifier along with Bayesian Networks (BN) classifier. The results of two classifiers are combined by taking an average of the probability distributions calculated by each of the two classifiers. Feature extraction and selection techniques have been incorporated with the model to find the most discriminating terms for classification. This classification model has been tested on three real text datasets. According to experiments, this approach showed better performance and the overall accuracy is higher than the accuracies of the two constituent classifiers. This technique also surpasses the accuracy of other well known, standard classifiers. This approach differs from the previous classification techniques in that it successfully incorporates MNB and BN classifiers and shows significantly better results than using either of the two classifiers separately. A comparative study of previous approaches with our method indicates a significant improvement over a number of techniques that were evaluated on the same dataset.

Marcos Mouriño et. al. [2] reports a comprehensive experimental evaluation of the efficiency of a bag-of-concepts representation for Bayesian text classification, tackling synonymy and polysemy, and exploiting semantic relatedness between concepts to alleviate the problem of orthogonality. Results of experiments performed on three corpora widely used as benchmarks—Reuters, OHSUMED, and 20 Newsgroups—show that: the efficiency of the bag-of-concepts approach is very dependent on the capacity of the semantic

annotator for extracting concepts and on the characteristics of particular corpora, peaking on OHSUMED; and that it performs especially well when the number of training samples is small. In particular, for the shorter training sequence bag-of-concepts outperforms bag-of-words by 3.67% in OHSUMED and 22.44% in 20Newsgroups.

Bo Tang et. al. [3] presenting a Bayesian classification approach for automatic text categorization using class-specific features. Unlike the conventional approaches for text categorization, our proposed method selects a specific feature subset for each class. To apply these class-dependent features for classification, we follow Baggenstoss's PDF Projection Theorem to reconstruct PDFs in raw data space from the class-specific PDFs in low-dimensional feature space, and build a Bayes classification rule. One noticeable significance of our approach is that most feature selection criteria, such as Information Gain (IG) and Maximum Discrimination (MD), can be easily incorporated into our approach. We evaluate our method's classification performance on several real-world benchmark data sets, compared with the state-of-the-art feature selection approaches. The superior results demonstrate the effectiveness of the proposed approach and further indicate its wide potential applications in text categorization.

Yin Aphinyanaphongs et. al. [4] This feasibility study explores text classification methodologies for identifying alcohol use tweets. We labeled 34,563 geo-located New York City tweets collected in a 24 hour period over New Year's Day 2012. We preprocessed the tweets into stem/ not stemmed and unigram/ bigram representations. We then applied multinomial naïve Bayes, a linear SVM, Bayesian logistic regression, and random forests to the classification task. Using 10 fold cross-validation, the algorithms performed with area under the receiver operating curve of 0.66, 0.91, 0.93, and 0.94 respectively. We also compare to a human constructed Boolean search for the same tweets and the text classification method is competitive with this hand crafted search. In conclusion, we show that the task of automatically identifying alcohol related tweets is highly feasible and paves the way for future research to improve these classifiers.

Richard McAllister et. al. [5] presents a new method for preparing a dataset for probabilistic classification by determining, a priori, the utility of a very small subset of taxonomically related dimensions via a Discriminative Multinomial Naiıve Bayes process. Author show that this method yields significant improvements over both Discriminative Multinomial Naiıve Bayes and Bayesian network classifiers alone.

## 3. OBJECTIVES

a. To discover new knowledge within text collections

b. Identification of similarities between text attributes that originate from different sources.

c. Reducing redundant information and matching same entities across different

d. Sources and various representations.

## 4. PROPOSED ALGORITHM

Naive Bayes classifier has been widely used for text categorization due to its simplicity and efficiency. It is a model-based classification method and offers competitive MNB would be one of the best-known naive Bayes classification approaches using the term frequency to represent the document. Considering a text categorization problem with N classes ("topics"), let c be the discrete variable of class label taking values in f1; 2; . . .;Ng, and x be the integer- valued feature vector corresponding to the term frequency. The MNB classifier assumes that the number of times that each term appears in the document satisfies a multinomial distribution [13], [20]. In other words, a document with l terms is considered as l independent trials, and each term is the result of a trial exactly falling into the vocabulary. Let the vocabulary size be M, and then each document is represented by a M _ 1 feature vector. Hence, given a document D, we first count the number of times that each term appears and generate a feature vector d ¼ ½x1; x2; . . . ; xM_T . According to the multinomial distribution, the likelihood of observing d conditioned on the class label c and the document length l can be calculated as follows:

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}$$

Where,

C is a hypothesis,

P (C) denotes the prior probability of C,

*P*(A) denotes the prior likelihood of the training data A being observed (i.e., given no knowledge about which hypothesis holds), and P(A|C) denotes the likelihood of observing data A given hypothesis C.

The left side of the formula P(C|A) is called posterior probability of C given the observed data *A*, which is the probability that we usually intend to obtain in a machine learning problem. Hence, the Bayes' theorem provides us a way to calculate the posterior probability P(C|A) from the data likelihoods P(A) and P(A|C), as well as the prior probability P(C).

## 5. SYSTEM ARCHITECTURE

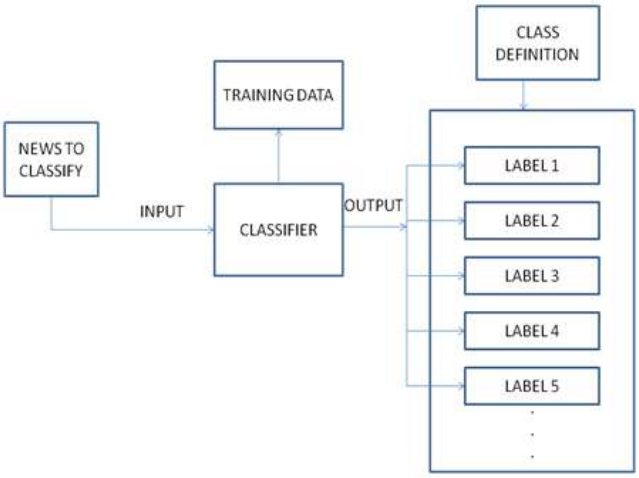


**Figure 1:** System Architecture

From the above figure, you can see the system architecture clearly. System consists of new to be classified as its input and the output will be the label to which the news probably belongs to. Classifier is the main module of the system which is the implementation of the naïve bayes algorithm. It uses the training data as its input and classifies the input documents. Training data consist of large number of documents preprocessed i.e. term frequency and document frequency is calculated. Using this data the input file is classified.

## 6. RESULT AND DISCUSSION

This section presents the screenshots of the working system in order to demonstrate the complete process of the system. The first screen, after starting the system shown to the user is displayed below in the screenshot, Figure .


**Figure 2:** Selection Page


**Figure 3:** selection of text

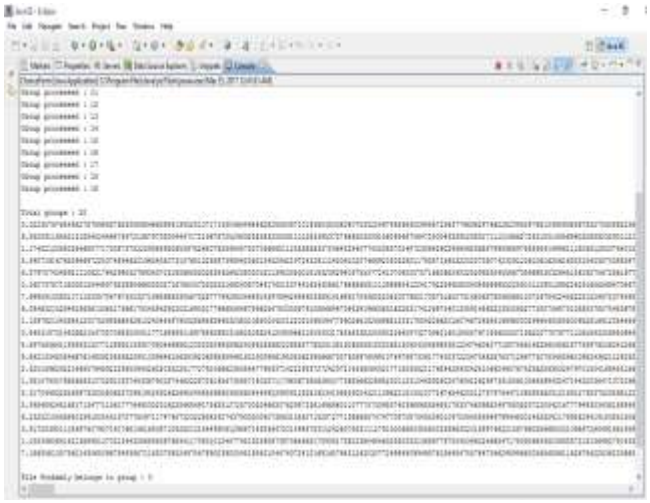

**Figure 4:** Text Classification

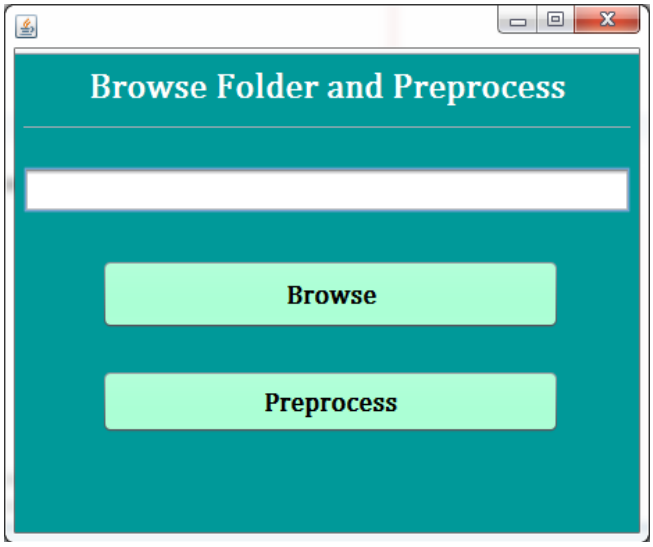


**Figure 4:** Deign Preview Second Page

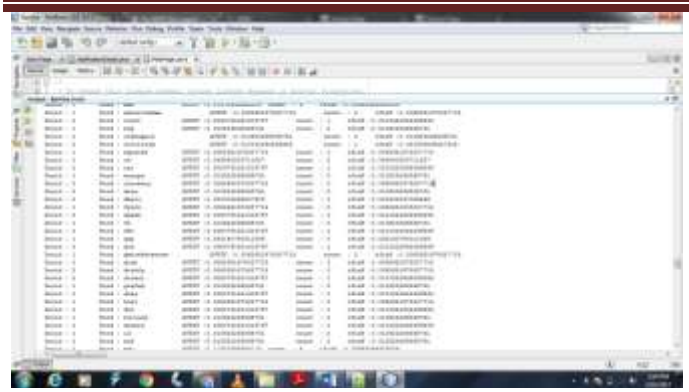

**Figure 5:** Text Reduction Fourth Page


**Figure 6:** Threshold

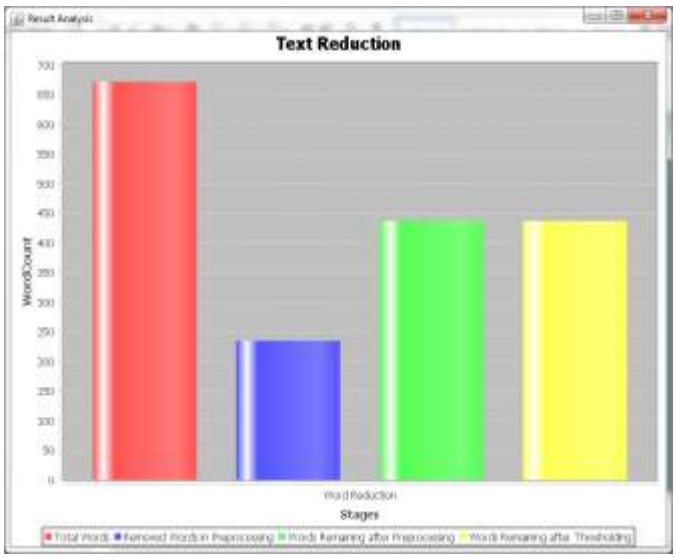


**Figure 7:** Text Reduction Graph Page

| Program | Icsiboost-bigram | Expected Maximum algorithm | Naïve Bayes Classifier |
|---|---|---|---|
| **Accuracy** | 71.6% | 79% | 86% |

Table1: Comparative analysis

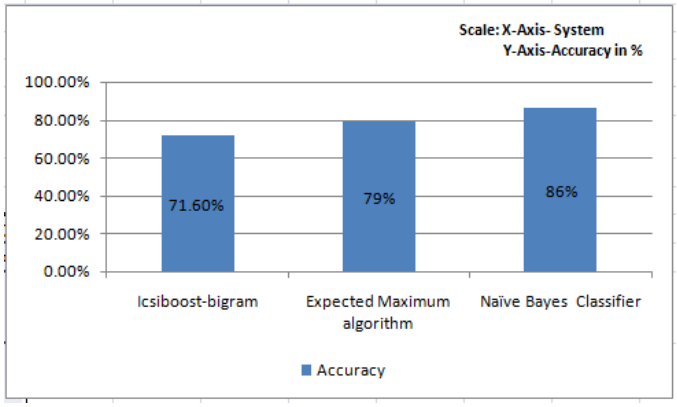


Figure 1 : Comparative analysis Accuracy Graph

Table 1 presents the accuracy of mentioned studies and compares the results with Naïve Bayes Classifier where its accuracy is obtained with this study.

## 7. CONCLUSION

Thus, The Text Classification using analytical approach project proposed a design of the application that can effectively classify text files into appropriate folder depending upon the theme of the file, using the training data to model the classifier. This application automates the text classification process otherwise would take long time doing manually the same task. Text file are appropriately classified using this application. This application allows you to select the test data, training data. In the future, a similar concept can be used for different purposes like arrange your computer, classify various documents with various applications and analyze them.

## 8. REFERENCES

[1] Amna Rahman, Usman Qamar, "A Bayesian Classifiers based Combination Model for Automatic Text Classification",IEEE, 978-1-4673-9904-3/16/$31.00 ©2016

[2] Marcos Mouriño-García, Roberto Pérez-Rodríguez, Luis Anido-Rifón, Miguel Gómez-Carballa, "Bag-of-Concepts Document Representation for Bayesian Text Classification", Bag-of-Concepts Document Representation for Bayesian Text Classification, 2016.

[3] Bo Tang, Haibo He, Paul M. Baggenstoss, and Steven Kay,"A Bayesian Classification Approach Using Class-Specific Features for Text Categorization", IEEE, Transactions on Knowledge and Data Engineering, 2016.

[4] Yin Aphinyanaphongs, Bisakha Ray, Alexander Statnikov, Paul Krebs, "Text Classification for Automatic Detection of Alcohol Use-Related Tweets A Feasibility Study", IEEE IRI, San Francisco, California, USA, August 13-15, 2014.

[5] Richard McAllister, John Sheppard, Rafal Angryk, "Taxonomic Dimensionality Reduction in Bayesian Text Classification", 11th International Conference on Machine Learning and Applications 2012.

[6] Tushar Ghorpade, Lata Ragha, "Featured Based Sentiment Classification for Hotel Reviews using NLP and Bayesian Classification", International Conference on Communication, Information & Computing Technology (ICCICT), Oct. 19-20, Mumbai, India, 2012.

[7] Anagha R Kulkarnia, Vrinda Tokekarb, Parag Kulkarnic, "Identifying Context of Text Documents using Naïve Bayes Classification and Apriori Association Rule Mining"

[8] Aaditya Jain,Jyoti Mandowara,"Text Classification by Combining Text Classifiers to Improve the Efficiency of Classification",International Journal of Computer Application (2250-1797),Volume 6– No.2, March- April 2016

[9] Kapila Rani,Satvika,"Text Categorization on Multiple Languages Based On Classification Technique", International Journal of Computer Science and Information Technologies,Vol.7(3),2016.

[10]Mr. Prashant G. Ghulaxe,Dr. Anjali B. Raut,"Text Clustering and Classification: A Review",International Conference on Recent Trends in Engineering Science and Technology, January 2017