# Text Classification for Student Data Set using Naive Bayes Classifier and KNN Classifier

Rajeswari R.P[#1], Kavitha Juliet[#2], Dr.Aradhana[#3]

[#1] *Asst Prof, Dept of CSE, RYMEC, Ballari.*

[#2] *Asst Prof, Dept of CSE, RYMEC, Ballari.*

[#3] *Professor, Dept of CSE, BITM, Ballari.*

***Abstract*** *— In this Information Era, Text documents with large features are available in plenty. Correct classification of this text documents into predefined set is a critical task. Text document classification is an emerging field in the area of text mining. Text classification is gorgeous because it eliminates the need of manually organizing documents based on their content and provides good accuracy. For Automated Text Classification a number of classifiers are available.* ***I****n this paper, our focus is on text classification using Naïve Bayes classifier and K-Nearest Neighbour classifier and to emphasize on performance and accuracy of these classifiers using Rapid miner for Student Data Set.*

***Keywords:*** *Text Mining, Text Classification, Naïve Bayes Classifier, KNN Classifier****.***

## I. INTRODUCTION

As the size of text documents from various sources such as social media, web resources, Research papers, corporate documents and patient health care documents are increasing day by day, proper classification of these voluminous documents into predefined category is essential. We need to have well defined methodology to analyse and classify this data and get the useful insight. Text mining aims to facilitate users to extract constructive information from these textual resources .It deals with operations like retrieval, classification, clustering etc. Text classification is the process of categorizing documents into predefined class based on their content. The application of text classification includes email spam filtering, indexing of scientific articles, filing patents into patent directories and proper identification of document kind.

## II. TEXT CLASSIFICATION PROCESS

The task of Text Classification is carried out in several sub phases. They are Data Collection & Representation, Document Pre-processing phase,

Feature selection or Transformation, Applying a text classifier and performance evaluation.

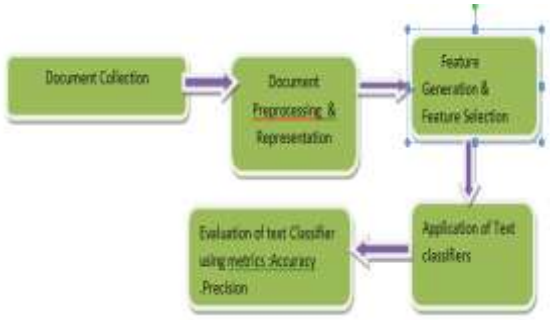


Fig 1: Text Classification

### *2.1 Document Collection*

This is first step of classification process. The text documents from various sources are collected in different formats of document like html,.doc, .pdf, web content etc.[1] . These documents are used during training and testing the classifier.

### *2.2 Document Pre-Processing &Representation*

A document is considered as collection of words or Bag of words [2]. Only few of these words are used for the classification .The words which are insignificant seems to degrade the process of classification. Hence these documents needs to be pre processed. The steps of Pre –processing are:

i) *Tokenization:* A document is defined as collection of strings of sentences, which are then isolated into set of tokens by removing spaces and commas [1]

*ii) Removing un useful stop words*: Certain words that are not useful for classification is called stop words. These are called Stop words. This step involves removal of frequently occurring Stop words like “the”, “a”, “and”, ’of”’, etc .from the list of tokens.

*iii) Stemming word:* Another very important step to reduce the number of words is to use stemming which converts different word form into similar canonical form. This technique is used to find the root or stem of a word. Stemming converts words to their root words. For example the words like waiting, raining converted to wait, rain respectively

***2.3 Feature Generation/Text Transformation****:-* Text Transformation or Feature generation is one of important step of pre-processing, which reduces the

complexity of documents. Documents have to be transformed from the full text version to a vector of words using the vector Space model.

### *2.4 Feature Selection*

Feature Selection is the dimension reduction step. The main aim of Feature selection is to select subset of features from the original Documents. Feature Selection is performed by retaining the most significant words according to predetermined measure of the significance of the word. Most important feature evaluation metric are information gain (IG), Chi-square, expected cross entropy, term frequency, Odds Ratio, the weight of evidence of text, mutual information, Gini index. Advantage of Feature selection is that it improves the accuracy and efficiency of the classifier .Also feature selection reduces the size of dataset and provides minimum search space.

### *2.5 Text Classification*

The objective of the classification is to classify the pre processed documents into predefined categories by using the training data set. The documents can be classified by three ways: supervised, unsupervised and semi supervised methods. In Supervised: All data is labelled and the algorithms learn to predict the output from the input data. Unsupervised: All data is unlabeled and the algorithms learn to inherent structure from the input data. Semi-supervised: Some data is labelled but most of it is unlabeled and a mixture of supervised and unsupervised techniques can be used. Classification is the problem of Supervised Learning. A number of Classifiers such as Bayesian classifier, Decision Tree, K-nearest neighbour (KNN), Support Vector Machines (SVMs), Neural Networks, are available.

### 2.6 Performance Evaluation

This is the last step of text classification in which text classifiers are evaluated as a measure of performance. , To measure the classifier precision, recall, and accuracy are most often used. *Precision wrt ci* is defined as the as the probability that if a random document *dx* is classified under *ci*, this decision is correct, *Recall wrt ci* is defined as the conditional that, if a random document *dx* ought to be classified under *ci*, this decision is taken

TPi–number of document correctly assigned to this category.

TNi - The number of document correctly rejected assigned to this category

FPi - The number of document incorrectly rejected assigned to this category

FNi - number of document incorrectly assigned to this category

Precision= TPi/TPi +FPi

Recall= TPi/TPi +FNi

Accuracy = TPi + TNi [1].

## *III LITERATURE SURVEY*

[1].Bhumika, Prof Sukhjit Singh Sehra, et al (2013) discusses the various Text classification algorithms that can be used for classification. Text classification is the task of categorizing a set of documents into predefined set. Text classification model such as Decision tree., Neural network and Generic algorithm are discussed.

[2].Vandana Korde et al (2012) discusses the importance of Text classification and several existing classification methods are discussed and compared with other classification methods based on few parameters such as classification criteria and time complexeties

[3].Ikonomakis, M.,S. Kotsiantis, and V. Tam pakas .discusses that Automated text classification is a crucial method to manage and process a enormous amount of text documents in digital forms .Few of the Machine learning algorithms such as Naives Bayes, Support Vector machine, Neural networks ,Nearest Neighbour and decision tress are discussed.

[4]Kamruzzaman, S. M., Farhana Haider, and Ahmed Ryadh Hasan.discusses that Text classification plays vital role for retrival systems. Several of the most common techniques used for text classification such as Association Rule Mining, Implementation of Naïve Bayes Classifier, Genetic Algorithm, Decision Tree are discussed.

[5] Menaka, S., and N. Radha. discusses that Text classification is the process of classifying the text documents based on words, phrases and word combinations with respect to set of predefined categories..Keywords are subset of words that contains the most important information about the content of the document. In the proposed system keywords are extracted from documents using TF-IDF and WordNet..The experiment has been carried for Naive Bayes, Decision tree and K-Nearest Neighbor (KNN) algorithms and its performance s are analyzed. Decision tree algorithm gives the better accuracy for text classification when compared to other algorithms.

## IV CLASSIFICATION ALGORITHMS

### *4.1 KNN classifier*

KNN classifier is case based machine learning algorithm [11] used for the automatic classification or categorization of text documents.KNN classifier is based on the measure of Euclidean distance or measure of similarity between documents and k training data. This Classifier emphasizes on the measure of similarity for identifying neighbours of particular document.KNN is easy to implement, it is

effective and non parametric. The drawback of KNN is its long time taken for classification

### *4.2 Naïve Bayes*

Navie Bayes classifier is simple classifier with is based on Bayes Theorem of conditional probability and strong independence assumptions. This classifier emphasizes on measure of probability that whether the document A belongs to class B or not. It is based on independent feature model. It is based on the assumption that occurrence or non occurrence of a particular attribute is unrelated to the occurrence or non occurrence of a particular attribute [4]. The advantage of Bayesian classifier is that it requires small training data set for classification. It is easier for implementation, fast to classify and more efficient .It is non sensitive to irrelevant features. It is used in personal email sorting, document categorization, email spam detection and sentiment detection [10]

## *V EXPERIMENTATION & RESULTS*

In this paper, experiment is carried out using Rapid miner tool 5.3.we are taking into consideration the student excel sheet with the following attributes. <USN, Course name, College name, Sex, Age, Year of study, AGPY, Scholarship >.The Design perspective is as shown in the fig 3.First ,Retrieve operator is used to access the data repository which contains Student Excel sheet. This example set is fed as input to select attribute operator. Select attribute operator selects which attributes of an *Example Set* should be kept and which are to be removed. Year of Study, AGPY and Scholarship has been selected

### *5.1 Working of Naive Bayes*

The Retrieve operator is used to load the 'Student scholarship' data set. The Select Attributes operator is applied on it to select just, AGPY, Scholarship attributes. The Naive Bayes operator is applied on it and the resulting model is applied on the 'Student scholarship' data set. The Same two attributes of the 'Student scholarship' data set were selected before application of the Naive Bayes model. The distribution table is as shown in the figure 4. Naive Bayes does calculation for all possible label values and selects the label value that has maximum calculated probability. The naïve bayes classifier shows that 8 out of 15 has label Yes .The probability of label = yes is 8/18. The probability of label = no is 4/18.The Distribution for class Y is 0.556 and for class N is 0.444 as shown in fig 4The performance vector shows that navies Bayes has accuracy of 66.7%.

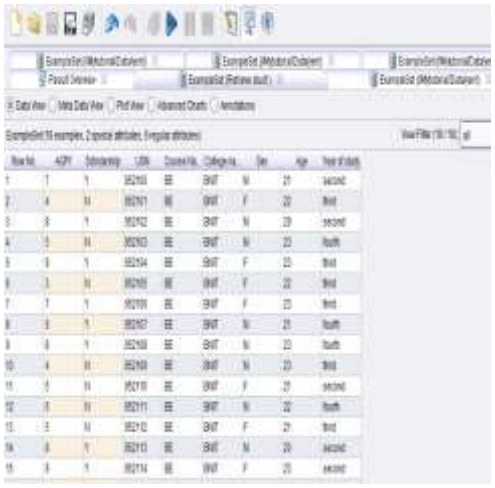

Figure 2: Example data Set

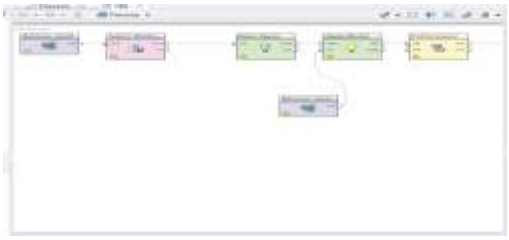

Figure 3: Design Perspective after applying Naïve Bayes.


Figure 4: Distribution table for Naives Classifier.


Figure 5: Performance Vector

### *5.2 Working of KNN.*

The k-Nearest Neighbour algorithm classifies by comparing a given test example with training examples that are similar to it. The training examples are described by n attributes. Each example represents a point in an n-dimensional space. All of the training examples are stored in an n-dimensional pattern space. When given an unknown example, a k-nearest neighbour algorithm searches the pattern space for the k training examples that are closest to the unknown example. These k training examples are the k "nearest neighbours" of the unknown example. "Closeness" is defined in terms of a distance metric, such as the Euclidean distance. This classifier gives an accuracy of 38.89% as shown in the fig 7. Further the graph in fig 9 shows the accuracies of the two classifier.


Figure 6: KNN Classification

Figure 7: Accuracy table for KNN

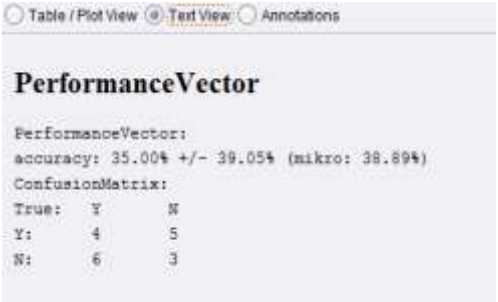


Figure 8: Performance Vector.

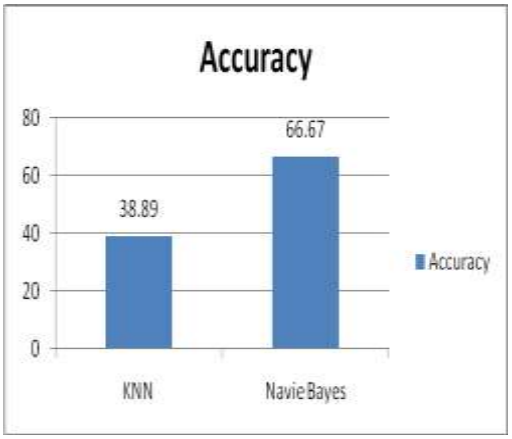


Figure 9: Accuracy of Naïve and KNN

## *VI CONCLUSION*

Text classification is one of the important phases of text mining, Text classification is a widespread task of classifying documents into predefined categories based on their contents

Application areas of text or document classification are opinion mining, sentiment analysis, labelling, text object recognition from news document etc. Naives Bayes Classifier and KNN classifiers are applied to student data set. The experiment carried out shows that Naives Bayes classifier is good classifier with accuracy of 66.67 than KNN classifier with 38.89.

## References

[1].Bhumika, Prof Sukhjit Singh Sehra, and Prof Anand Nayyar. "A review paper on algorithms used for text classification." *International Journal of Application or Innovation in Engineering & Management* 3.2 (2013): 90-99.

[2]. Korde, Vandana, and C. Namrata Mahender. "Text classification and classifiers: A survey." *International Journal of Artificial Intelligence & Applications* 3.2 (2012): 85.

[3].Ikonomakis, M., S. Kotsiantis, and V. Tampakas. "Text classification using machine learning techni ques." *WSEAS transactions on computers* 4.8 (2005): 966-974.

[4] Kamruzzaman, S. M., Farhana Haider, and Ahmed Ryadh Hasan. "Text classification using data mining." *arXiv preprint arXiv:1009.4987* (2010).

[5] Joachims, Thorsten. "Text categorization with support vector machines: Learning with many relevant features." *European conference on machine learning*. Springer Berlin Heidelberg, 1998.

[6]. Menaka, S., and N. Radha. "Text classification using keyword extraction technique." *International Journal of Advanced Research in Computer Science and Software Engineering* 3.12 (2013).

[7] Williamson, Eric R., and Saurabh Chakravarty. "CS5604 Fall 2016 Classification Team Final Report." (2016).

[8] Dalal, Mita K., and Mukesh A. Zaveri. "Automatic text classification: a technical review." *International Journal of Computer Applications* 28.2 (2011): 37-40.

[9] Ting, S. L., W. H. Ip, and Albert HC Tsang. "Is Naive Bayes a good classifier for document classification?." *International*

*Journal of Software Engineering and Its Applications* 5.3 (2011): 37-46.

[10] Mahesh Kini M , Saroja Devi H , Prashant G Desai, Niranjan Chiplunkar." Text Mining Approach to Classify Technical Research Documents using Naïve Bayes" *International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 7, July 2015*

[11]. Gongde Guo, Hui Wang, David Bell, Yaxin Bi and Kieran Greer, "KNN Model-Based Approach in Classification", Proc. ODBASE pp- 986 – 996, 2003