



# INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

Impact factor: 4.295

(Volume2, Issue6)

Available online at: [www.Ijariit.com](http://www.Ijariit.com)

## Data Mining: Text Classification System for Classifying Abstracts of Research Papers

Shirdi Wazeed Baba<sup>1</sup>, Reddi Sanjeev Kumar<sup>2</sup>

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, DESIGN AND MANUFACTURING, JABALPUR.

[wazeed.7@gmail.com](mailto:wazeed.7@gmail.com)<sup>1</sup>, [reddisanjeevkumar@gmail.com](mailto:reddisanjeevkumar@gmail.com)<sup>2</sup>

---

*Abstract- Text classification is the process of classifying documents into predefined categories based on their content. Text classification is the primary requirement of text retrieval systems, which retrieve texts in response to a user query, and text understanding systems, which transform text in some way such as producing summaries, answering questions or extracting data. We have proposed a Text Classification system for classifying abstract of different research papers. In this System we have extracted keywords using Porter Stemmer and Tokenizer. The word set is formed from the derived keywords using Association Rule and Apriori algorithm. The Probability of the word set is calculated using naive bayes classifier and then the new abstract inserted by the user is classified as belonging to one of the various classes. The accuracy of the system is found satisfactory. It requires less training data as compared to other classification system.*

*Index Terms- Apriori algorithm, Association rule, confidence, support, Naïve Bayes classifier, Text classification.*

---

### I. INTRODUCTION

Constructing fast and accurate classifiers for large data sets is an important task in data mining and knowledge discovery. There are numerous text documents available in electronic form. More and more are becoming available every day. Such documents represent a massive amount of information that is easily accessible. Seeking value in this huge collection requires organization; much of the work of organizing documents can be automated through data mining. The accuracy and our understanding of such systems greatly influence their usefulness. There is growing evidence that merging porter stemmer, naive bayes and association rule mining together can produce more efficient and accurate classification systems than traditional classification techniques [26]. Classification is one of the most important tasks in data mining. There are many classification approaches for extracting knowledge from data such as statistical [21], divide-and-conquer [15] and covering [6] approaches. Numerous algorithms have been derived from these approaches such as genetic algorithm. However, traditional classification techniques often produce a small subset of rules, and therefore usually miss detailed rules that might play an important role in some cases [29]. The task of data mining is to automatically classify documents into predefined classes based on their content. Many algorithms have been developed to deal with automatic text classification [4]. With the existing algorithms, a number of newly established processes are involving in the automation of text classification [20]. The most common techniques used for Text Classification include Association Rule Mining, Naïve Bayes Classifier, Decision Tree and others. Association rule mining finds interesting association or correlation relationships among a large set of data items [1][4]. The discovery of these relationships among huge amounts of transaction records can help in many decision making process. On the other hand, the Naïve Bayes classifier uses the maximum a posteriori estimation for learning a classifier. It assumes that the occurrence of each word in a document is conditionally

independent of all other words in that document given its class [3]. Each abstract is considered as a transaction in the text data. After pre-processing the text data association rule mining [1] is applied to the set of transaction data where each frequent word set from each abstract is considered as a single transaction. This paper presents a new algorithm for text classification.

Porter stemmer algorithm is used to remove unnecessary words from abstracts. The association rule is used to derive feature sets from pre-classified text documents. The concept of Naive Bayes Classifier is then used on derived features sets to calculate the probability of derived word sets.

The paper is structured as follows: Section 2 describes the previous work already done in the field of text classification, section 3 describes the concept of text classification, section 4 contains the proposed algorithm, and section 5 contains the experimental results defining the datasets used and various abstracts taken for classifying the text.

## **II. PREVIOUS WORK**

Classification is to put things according to their characteristics. Given a set of class, classifier determines which classes a given object belongs to. Documents may be classified according to their subjects or the other attributes such as document type, author, printing year etc. In Text Classification the most popularly used approaches are Porter Stemmer, Apriori algorithm, Naïve Bayes etc. Most of the researches in text categorization come from the machine learning and information retrieval communities such as decision trees, Naïve Bayes (NB)[9], Support Vector Machines (SVM) [11], k-Nearest Neighbor (kNN) [10], Neural Network (NNet) and etc. Among these methods, KNN is a simple statistic method and it also shows good performance. The automatic classification of documents into predefined categories can be classified by three ways: unsupervised, supervised, and semi-supervised methods [11]. From last few years, the task of automatic text classification has been extensively studied and rapid progress seems in this area, including the machine learning approaches. Vandana Korde et al (2012) [21] observed that the text mining studies are gaining more importance recently because of the availability of the increasing number of the electronic documents from a variety of sources which include unstructured and semi structured information. The main goal of text mining is to enable users to extract information from textual resources and deals with the operations like, retrieval, classification(supervised, unsupervised and semi supervised) and summarization, Natural Language Processing (NLP), Data Mining, and Machine Learning techniques work together to automatically classify and discover patterns from the different types of the documents.

Zakaria Elberrichi, et al (2008) [21] presents a new approach for text categorization based on incorporating background knowledge (Word Net) into text representation with using the multivariate, which consists of extracting the K better features characterizing best the category compared to the others. Newsgroups datasets show that incorporating background knowledge in order to capture relationships between words is especially effective in raising the macro-averaged F1 value. William B. Cavnar et al (2010) [21] proposed a N-gram frequency method that provides an inexpensive and highly effective way of classifying documents. It does so by using samples of the desired categories rather than resorting to more complicated and costly methods such as natural language parsing or assembling detailed lexicons. Essentially this approach defines a “categorization by example” method. Collecting samples and building profiles can even be handled in a largely automatic way. Also, this system is resistant to various OCR problems, since it depends on the statistical properties of N-gram occurrences and not on any particular occurrence of a word.

Andrew McCallum [21], has compared the theory and practice of two different first-order probabilistic classifiers, both of which make the naive Bayes assumption." The multinomial

model is found to be almost uniformly better than the multi variant Bernoulli model. In empirical results on five real-world corporate.

Author Mitchell [21], used training data for learning to classify text from all three categories, of which 47 are from Computer Science, 48 are from Electrical and Electronic Engineering and the rest 20 are from Mechanical Engineering papers. After preprocessing the text data association rule mining is applied to the set of transaction data where each frequent word set from each abstract is considered as a single transaction.

## **III. BACKGROUND STUDY**

### **3.1 Data Mining**

Data Mining is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data. The information or knowledge extracted so can be used for any of the following applications: Market Analysis, Fraud Detection, Customer Retention, and Production Control Science Exploration.

### **3.2 Classification**

Classification has two distinct meanings. We may be given a set of observations with the aim of establishing the existence of classes or clusters in the data. Or we may know for certain that there are so many classes, and the aim is to establish a rule whereby we can classify a new observation into one of the existing classes. The former type is known as Unsupervised Learning (or Clustering), the latter as Supervised Learning. In the statistical literature, Supervised Learning is usually, but not always, referred to as discrimination, by which is meant the establishing of the classification rule from given correctly classified data.

### 3.5 Naive Bayes Classifier

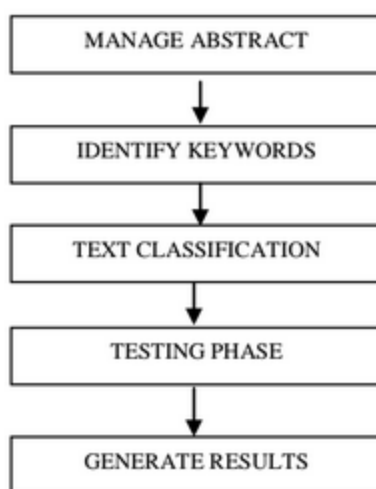
The Naive Bayes algorithm is a classification algorithm based on Bayes rule and a set of conditional independence assumptions. Given the goal of learning  $P(Y|X)$  where  $X = \{X_1, \dots, X_n\}$ , the Naive Bayes algorithm makes the assumption that each  $X_i$  is conditionally independent of each of the other  $X_k$ s given  $Y$ , and also independent of each subset of the other  $X$ 's given  $Y$ . The value of this assumption is that it dramatically simplifies the representation of  $P(X|Y)$ , and the problem of estimating it from the training data. Consider, for example, the case where  $X = \{X_1, X_2\}$ . In this case.

$$\begin{aligned} P(X|Y) &= P(X_1, X_2|Y) \\ &= P(X_1|X_2, Y)P(X_2|Y) \\ &= P(X_1|Y)P(X_2|Y) \end{aligned}$$

Where the second line follows from a general property of probabilities, and the third line follows directly from our above definition of conditional independence. More generally, when  $X$  contains  $n$  attributes which satisfy the conditional independence assumption, we have  $P(X_1 \dots X_n|Y) = \prod_{i=1}^n P(X_i | Y)$ .

## IV. PROPOSED WORK

In this work a text classification system is proposed. Our method to classify text is an implementation of Porter Stemmer with a combined use of Naïve Bayes Classifier and Association Rule. We have used the features of association rule to make association sets. On the other hand, to make a probability chart with prior probabilities we have used Naïve Bayes classifier's probability measurements. And in the last retrieval phase of test data we have implemented the positive-negative matching calculation observed in different researches [2][6]. Here the associated word sets, which do not match our considered class is treated as negative sets and others are positive. Flowchart for the proposed method is given in Figure 4.1.



**Figure 4.1: Proposed Method**

4.1 The Algorithm for Text Classification The proposed algorithm uses various steps for classifying text; these steps are described below in detail.

i) Porter Stemmer To make the raw text valuable, that is to prepare the text, we have considered only the keywords. That is unnecessary words and symbols are removed. For this keyword extraction process we dropped the common unnecessary words like am, is, are, to, from...etc. and also dropped all kinds of punctuations and stop words. Singular and plural form of a word is considered same. Finally, the remaining frequent words are considered as keywords. The text data is cleaned by removing unnecessary words i.e. text data is filtered and subject related words are collected.

Input: Database, D Minimum support threshold, min\_sup. Output: L, frequent itemsets in D.

#### ii) The Apriori Algorithm

Keywords obtained from Porter Stemmer are joined together to form word sets. Each frequent word set from each abstract is considered as a single transaction. Using these transactions, we generated a list of maximum length sets applying the Apriori algorithm. The Apriori algorithm is given below:

Input: Database, D; Minimum support threshold, min\_sup.

Output: L, frequent item sets in D.

#### iii) Association Rule

For each frequent word set obtained from Apriori confidence and support is calculated in Association Rule Mining. Association rule mining finds interesting association or correlation relationships among a large set of data items. The discovery of these relationships among huge amounts of transaction records can help in many decision making process. In this project association rules from the significant words derived from keyword extraction apriori algorithm are used to derive feature set from pre-classified text documents.

iv) Naive Bayes Classification It calculates the probability of different class with the probability values of the matched set obtained from association rule mining while ignoring the unmatched sets. As a result set if test set matches with a rule set, which has weak probability to the actual class, may cause wrong classification. To make a probability chart with prior probabilities we have used Naïve Bayes classifier's probability measurements.

The algorithm is as follows:

1. For each class  $i = 1$  to  $n$  do
2. Set  $pval = 0$ ,  $nval = 0$ ,  $p = 0$ ,  $n = 0$
3. For each set  $s = 1$  to  $m$  do
4. If the probability of the class (i) for the set (s) is maximum then increment  $pval$  else increment  $nval$ .
5. If 50% of the associated set  $s$  is matched with the keywords set do step 6 else do step 7
6. If maximum probability matches the class  $i$  then increment  $p$
7. If maximum probability does not match the class increment  $n$
8. If  $(s \leq m)$  go to step 3
9. Calculate the percentage of matching in positive sets for the class  $i$
10. Calculate the percentage of not matching in negative sets for the class  $i$
11. Calculate the total probability as the summation of the results obtained from step 9 and 10 and also the prior probability of the class  $i$  in set  $s$
12. If  $(i \leq n)$  go to step 1
13. Set the class having the maximum probability value as the result

Where,  $n$  = number of class,  $m$  = number of associated sets,

$pval$  = positive value

$nval$  = negative value

$s$  = set

$i$  = increment variable

The experimentation work done is described in next section.

## V. EXPERIMENTAL WORK

### 5.1 Dataset

Abstracts from different research papers have been used as data sets for training and testing of the proposed method. Four classes of papers from DBMS, Operating System, Java and Data Structure were considered for our experiment. Total 40 abstracts are used in our experiment (10 from DBMS, 10 from Java, 10 from Operating System and 10 from Data Structure). The result obtained for training and testing is discussed in subsequent section.

### 5.2 Experimental Results

We have divided the abstract into training set and testing set and then proposed algorithm is used to perform our experiments. We have repeated the experiments by changing the abstracts in testing set and keeping all other parameter constant in training set. For Example abstract taken as input for Testing is: "With respect to all algorithm perspective coding binary trees and an representation for well-formed parentheses strings. We present here the first Gray code and loop less generating algorithm for P-sequences, and extend them in a Gray code and a new loop less generating algorithm for well-formed parentheses strings. Given a connected graph  $G = (V, E)$  and a spanning tree  $T$  of  $G$ , a fundamental cycle is a cycle resulting by adding an edge  $e \in E - T$  to  $T$ . In this paper we establish that the average length of fundamental cycles in a complete graph increases with the number of vertices. Also, given a simple cycle in a complete graph, the paper describes a method of calculating the number of spanning trees, with respect to which the cycle is a fundamental cycle." Output generated is: respect, algorithm, tree, well, formed, parenthesis, string, gray, code, loop less, generating, graph, fundamental, cycle, paper, complete, number. The keyword extraction process is applied to all the abstracts and the value calculated according to algorithm is as follows:

$pval=10, nval=30, p=2, n=30.$

Now the probability of DBMS =  $((p*100)/pval) + ((n*100)/nval)$

+ prior probability of DBMS

=  $((2*100)/10) + ((30*100)/30) + 0.26$

= 120.26

For this set of keywords,

Calculated Probability for class DBMS = 120.26

Calculated Probability for class Data Structure = 106.09

Calculated Probability for class Java = 118.08

Calculated Probability for class Operating System = 121.13

The results obtained by Proposed Method with Text Classifier for different abstracts are as shown in Table I.

Table I show results that are found using the same data sets for both Association Rule with Naïve Bayes Classifier and proposed method.

**Table I**

Technique	(%) Training Data	(%)Data Accuracy
Association Rule Based Decision Tree	76	87
Naïve Bayes Classifier	69	68
Proposed Algorithm	50	75

In text categorization using association rule based decision tree [16] 76 % data set of the total 40 data set was used to train and 87% accuracy was observed. On the other hand using 50% data as training data the proposed algorithm can able to classify text with 75% accuracy rate.

## VI. CONCLUSION

In this paper a Text Classification System for classifying abstracts of the research paper in four categories (Java, Operating System, DBMS, Data Structure) have been proposed. To improve the performance of classification, Association rule and Naïve bayes classifiers are used. To demonstrate and validate our approach we have presented the result on forty real datasets. To describe the usefulness of our approach we have compared the probability of various abstract papers and obtained the satisfactory results in term of accuracies. We have achieved 75% accuracy of classification.

## REFERENCES

- [1] Agarwal R., Mannila H., Srikant R., Toivonen H., Verkamo, "A Fast Discovery of Association Rules," Advances in knowledge Discovery and Data Mining, , pp. 138-189, 1996.
- [2] Anwar M. Hossain, Mamunur M. Rashid, Chowdhury Mofizur Rahman, "A New Genetic Algorithm Based Text Classifier," In Proceedings of International Conference on Computer and Information Technology, NSU, pp. 135-139, 2001.
- [3] Bing Liu, Wynne Hsu, Yiming Ma, "Integrating Classification and Association Rule Mining," In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98, Plenary Presentation), New York, USA, 1998.
- [4] Canasai Kruengkrai, Chuleerat Jaruskulchai, "A Parallel Learning Algorithm for Text Classification," The Eighth ACM International Conference on Knowledge Discovery and Data Mining (KDD-2002), Canada, July 2002, pp. 42-98.
- [5] Chowdhury Mofizur Rahman, Ferdous Ahmed Sohel, Parvez Naushad, Kamruzzaman S. M, "Text Classification Using the Concept of Association Rule of Data Mining," In Proceedings of International Conference on Information Technology, Nepal, pp 234-241, May 23-26, 2003.
- [6] Eshita Sharmin, Ayesha Akhter, Chowdhury Mofizur Rahman, "Genetic Algorithm for Text Categorization," In Proceedings of International Conference on Computer and Information Technology, BUET, pp. 80-85, December, 1998.
- [7] Han Jiawei, Micheline Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann Publisher: CA, 2001, pp. 20-55.
- [8] <http://www.pmsi.fr/gafxmpa.html>.
- [9] Jason D. M. Rennie, "Improving Multi-class Text Classification with Naive Bayes," 2001, Massachusetts Institute of Technology, <http://citeseer.ist.psu.edu/cs>.
- [10] Jason Kroll, "Decision Tree Learning for Arbitrary Text Classification," Sept 2003, [www.cs.tufts.edu/~jkroll/dectree](http://www.cs.tufts.edu/~jkroll/dectree) ICTM 2005
- [11] Kamruzzaman S. M, Farhana Haider, "A Hybrid Learning Algorithm for Text Classification", Accepted for the Publication of the Proceedings of the Third International Conference on Electrical and Computer Engineering, Going to be held at Dhaka, on December 28-30, 2004 , pp. 10-47, 2010.
- [12] Kamruzzaman S. M, Farhana Haider, Ahmed Ryadh Hasan, "Text Classification using Association Rule with a Hybrid Concept of Naive Bayes Classifier and Genetic Algorithm," Accepted for publication into International Conference on Computer and Information Technology (ICCIT-2004), Brac University, Dhaka, Bangladesh, to be held on December 26-28, 2004.
- [13] Lewis, D., and Ringuette, M., "A Comparison of Two Learning Algorithms for Text Categorization," In Third Annual Symposium on Document Analysis and Information Retrieval, pp. 81-93, 1994.
- [14] Loper Edward, "NLTK Tutorial: Text Classification," 2004, <http://nltk.sourceforge.net/tutorial>
- [15] Luger, George F, "Artificial Intelligence, Structures and Strategies for Complex Problem Solving," Fourth Edition, page 471.. 2002. Harlow, England.
- [16] Masud M. Hassan, Chowdhury Mofizur Rahman, "Text Categorization Using Association Rule Based Decision Tree," Proceedings of 6th International Conference on Computer and Information Technology, JU, pp.453-456, 2003.
- [17] McCallum, A., and Nigam, K., "A Comparison of Events Models for Naïve Bayes Text Classification," Papers from the AAI Workshop, pp. 41-48, 1998.
- [18] Sarah Zelikovitz and Haym Hirsh, "Integrating Background Knowledge into Nearest-Neighbor Text Classification," Proceedings of the 6th European Conference on Case Based Reasoning. Springer Verlag.
- [19] Yang Y., Zhang J. and Kisiel B, "A scalability analysis of classifiers in text categorization," ACM SIGIR'03, 2003, pp. 61- 110, 1999.
- [20] Bhumika<sup>1</sup>, Prof Sukhjot Singh Sehra<sup>2</sup>, Prof Anand Nayyar<sup>3</sup> "A REVIEW PAPER ON ALGORITHMS USED FOR TEXT CLASSIFICATION", International Journal of Application or Innovation in Engineering & Management (IJAIEM), pp 92.