# Text Classification using Data Mining

Prasad Khatate
Computer Engineering
Atharva College of Engineering
Mumbai, India

Mohammed Nadeem
Computer Engineering
Atharva College of Engineering
Mumbai, India

Sagar Jogale
Computer Engineering
Atharva College of Engineering
Mumbai, India

Prof. Satish Ranbhise
Computer Engineering
Atharva College of Engineering
Mumbai, India

***Abstract*— Text classification is the process of classifying text documents into predefined categories based on their content. It is an automated assignment of natural language texts to predefined categories. Text classification is an important requirement of text retrieval systems, where they retrieve texts in response to a query from user, and text understanding systems, which transform text in such way as producing summaries, classifying papers, answering questions or extracting data. All the existing supervised learning algorithms to automatically classify the text need sufficient documents to learn accurately. This paper proposes a new algorithm for text classification using data mining that requires fewer data for training. Instead of using words and word relation i.e. association rules from these words are used to derive feature set from classified text documents. The concept of the Naïve Bayes classifier is then used for final classification.**

***Keywords*— *Text Classification, Data Mining, Naïve Bayes Classifier***

## I. INTRODUCTION

### A. *Need:*

There are a number of text documents available on the Internet. And much more are becoming available every day. Such documents represent a tremendous amount of text information that is easily accessible. Seeking value in this huge data requires an organization; the work of organizing documents can be automated through data mining. The accuracy and understanding of such systems highly influence their usefulness. The task of data mining is to automatically classify documents into predefined classes based on their content. With the existing algorithms, a number of newly established processes are involving in the automation of text classification. The most common techniques used for the purpose of text classification include Association Rule Mining, Genetic Algorithm, Implementation of Naïve Bayes Classifier, and Decision Tree and so on. Association rule mining finds interesting correlation or association relationships among a large set of data items. Finding these relationships among huge amounts of data can help in many decision-making processes. On the other hand, the Naïve Bayes classifier uses the maximum posterior estimation for learning a classifier. It believes that the occurrence of each word in a text document is conditionally independent of all other words in that text document given its class. Although the Naïve Bayes works well in many studies, it requires a large number of training documents for learning accurately.

### B. *Basic Concept:*

Text classification is an important part of text mining, it is like manually building automatic Text Classification systems by means of knowledge engineering techniques. An example of such systems would be to automatically label each incoming news story with a topic like “sports”, “technology”, “art”, or “politics”. A data mining classification task starts with a training set of documents that are already labeled with a class C1, C2 [7].The task is to make a classification model which is able to assign the correct class to a new document. Text classification has two labeling forms as single-label and multi-label. Single label document belongs to only one class and the multi-label document may be belong to more than one class. In this paper, we consider only single label document classification.

## II. PREVIOUS WORK

Classification is to put things according to their characteristics. Given a set of classes, a text classifier determines which classes a given object belongs to. Text documents are classified according to their subjects or the other attributes such as keywords, document type, author, printing year etc. In Text Classification the most used approaches are Porter Stemmer, Genetic Algorithm, and Naïve Bayes etc. [7]. Most of the researches in text classification come from the machine learning systems and information retrieval communities such as decision trees, Naïve Bayes, Support Vector Machines, k-Nearest Neighbor, Neural Network and etc. Among these methods, k-Nearest Neighbor is a simple statistic method and it also performs well. The automatic classification of documents into predefined categories can be classified by three ways: semi-supervised, unsupervised, supervised methods.

Vandana Korde (2012) observed that the text mining is gaining more importance recently because of the availability of the increasing number of the electronic documents from a variety of sources which include unstructured data [1]. To enable users to extract information from textual resources and deal with the operations like retrieval, classification (unsupervised, supervised and semi-supervised) and Natural Language Processing (NLP), summarization, Machine Learning techniques and Data Mining work together to automatically classify text document and discover patterns from the different types of the text documents.

Mohamed Amine Bentaalah, Abdelattif Rahmoun, and Zakaria Elberrichi (2008) proposed a new approach for text categorization based on incorporating background knowledge (WordNet) into text representation by using the multivariate, which consists of extracting the K better features for characterizing best the category compared to the others [2].

Deepika Atre, Anuradha Purohit, Deepika Atre, Payal Jaswani, Priyanshi Asawara in the paper: "Text Classification in Data Mining" provide the use of Porter stemmer for Stemming, Apriori Algorithm to generate Frequent item set, Association rule to find correlation relationships among a large set of data items and Naïve Bayes for classification [7].

Akansha Garg in paper: "Web Page Classification using FP-Growth Algorithm" provide advantages offered by algorithms such as FP-Growth are partly gained from the ordering process, which reduces the overall processing time by allowing the most common items to be processed more efficiently [5].

In the paper: "Performance comparison of Apriori and FP-Growth algorithms in generating association rules" by Daniel Hunyadi provides a comparison of performance between the two algorithms [6].

## III. PROPOSED SYSTEM

In this work, a text classification system is proposed. Our system to classify text is an implementation of combined use of Naïve Bayes Classifier and Association Rule. We have used the characteristics of association rule mining to make association/ relationship sets. On the other hand, to make a probability chart with prior probabilities we have used Naïve Bayes classifier algorithm for probability measurements. And in the last retrieval phase of our system, we have implemented the positive-negative matching calculation observed in different researches. Here the associated word sets, which do not match our examined class is treated as negative sets and others are positive.
Flowchart for the proposed method is given in Figure 3.1

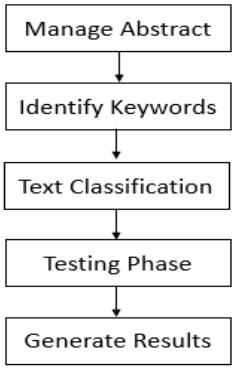


Fig. 6.1. Proposed System

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, we consider only the keywords i.e. unnecessary words and symbols are discarded. For this keyword extraction process, we dropped the common unnecessary words like am, is, are, to, from...etc. and we also discard all kinds of punctuations and stop words. The Singular and Plural form of a word are considered as the same word. Finally, the frequent words that remain are considered as keywords [7].The text is cleaned by removing unnecessary words i.e. text is filtered and subject-related words are collected.
Input: Database D, Minimum support threshold (min_sup)
Output: L, frequent item sets in D.

*ii) The FP-Growth Algorithm*
Keywords obtained from Porter Stemmer are joined together to form word sets. Each repeated word set from each abstract is considered as a single record. Using these records, we generated a record of maximum length sets applying the FP-Growth [5].The FP-Growth 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 FP-Growth, confidence and support is calculated in Association Rule Mining. Association rule mining finds captivating correlation or association relationships among a large set of data items. Finding these relationships among huge amounts of transaction records can help in many decision-making processes. In this project, association rules for significant words are derived from keyword extraction, FP-Growth algorithm is used to derive feature set from pre-classified text documents [7].

*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 a weak probability to the actual class, may cause the wrong classification. To make a probability table with prior probabilities we have used Naïve Bayes classifier's probability measurements. [7]

The algorithm is as follows:
1. For each class j = 1 to n do
2. Set px = 0, nx = 0, p = 0, n = 0
3. For each set s = 1 to m do
4. If the probability of the class (j) for the set (s) is maximum then
increment px
else
increment nx
5. If 50% of the associated set s is matched with the keywords
set do step 6 else do step7
6. If maximum probability matches the class j then increment p
7. If maximum probability does not match the class j then increment n
8. If (s<=m)
go to step 3
9. Calculate the percentage of matching in positive sets for the class j

10. Calculate the percentage of not matching in negative sets for the class j
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 j in set s
12. If (j<=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,*
*px =positive value , nx =negative value*
*s =set, j =increment variable*

## IV. CONCLUSION

Thus in this paper, we have focused on how to make the process of Text Classification more efficient. We have used various data mining techniques for the classification process. For more accuracy or optimality in classification, we have also used Association rule mining and Naive Bayesian algorithm which will help in minimizing the search time for classification category and system will be able to give optimized results.

## ACKNOWLEDGMENT

It gives us immense pleasure in presenting this project report titled: “Text Classification using Data Mining”. We wish to express our immense gratitude to the range of people who provided all the support needed for the completion of this project. Their guidance and encouragement have helped in making this project a great success.

We would like to express our sincere gratitude to our respected principal Dr. Shrikant Kallurkar and the organization of Atharva College of Engineering for providing such an ideal atmosphere to build up this project with a well-equipped library with all the utmost necessary reference materials and up to date Laboratories.

We are eager and glad to express our gratitude to the Head of Computer Engineering Department Prof. Mahendra Patil, for his approval of this project. We are also thankful to him for providing us the needed assistance, detailed suggestions and also the encouragement to do the project.

We express our gratitude to our project guide and mentor Prof. Satish Ranbhise, who provided us with all the guidance and encouragement and making the lab available to us at any time. We are extremely thankful to all other staff and the management of the college for providing us all the facilities and resources required.

## REFERENCES

[1] Vandana Korde , “Text Classification and Classifiers: Survey”, 2012.

[2] Abdelattif Rahmoun, Zakaria Elberrichi, and Mohamed Amine Bentaalah “Using WordNet for Text Categorization” The International Arab Journal of Information Technology, Vol.5, No. 1, January 2008

[3] Han Jiawei, Micheline Kamber, “Data Mining: Concepts and Techniques,” Morgan Kaufmann Publisher: CA, 2001, pp. 20-55.

[4] http://www.pmsi.fr/gafxmpa.html.

[5] Akansha Garg , “Web Page Classification using FP Growth Algorithm”, International Journal of Advanced Computer Technology (IJACT),ISSN:2319-7900.

[6] Daniel Hunyadi , Proceedings of the European Computing Conference, ISBN: 978-960-474-297-4 ,“Performance comparison of Apriori and FP-Growth algorithms in generating association rules”.

[7] Anuradha Purohit, Deepika Atre, Payal Jaswani, Priyanshi Asawara , “Text Classification in Data Mining”, International Journal of Scientific and Research Publications, Volume 5, Issue 6, June 2015, ISSN 2250-3153.