# COMPARATIVE STUDY OF CLASSIFICATION ALGORITHM FOR TEXT BASED CATEGORIZATION

**Omkar Ardhapure[1], Gayatri Patil[2], Disha Udani[3], Kamlesh Jetha[4]**

[1]*Student, Department of Computer Engineering, ABMSP's APCOER Pune, Maharashtra, India*
[2]*Student, Department of Computer Engineering, ABMSP's APCOER Pune, Maharashtra, India*
[3]*Student, Department of Computer Engineering, ABMSP's APCOER Pune, Maharashtra, India*
[4]*Assistant Professor, Department of Computer Engineering, ABMSP's APCOER Pune, Maharashtra, India*

## Abstract

*Text categorization is a process in data mining which assigns predefined categories to free-text documents using machine learning techniques. Any document in the form of text, image, music, etc. can be classified using some categorization techniques. It provides conceptual views of the collected documents and has important applications in the real world. Text based categorization is made use of for document classification with pattern recognition and machine learning. Advantages of a number of classification algorithms have been studied in this paper to classify documents. An example of these algorithms is: Naive Bayes' algorithm, K-Nearest Neighbor, Decision Tree etc. This paper presents a comparative study of advantages and disadvantages of the above mentioned classification algorithm.*

***Keywords:*** *Data Mining, Text Mining, Text Categorization, Machine Learning, Pattern Analysis, Naive Bayes', KNN, Decision Tree.*

----------***----------

## 1. INTRODUCTION

Text classification is an integral part of text mining for machine learning techniques. With the rapid growth of digital information, text categorization is widely being used to handle and organize text data. The main goal of text categorization is to divide free text documents into different categories and automatically assign documents into defined categories. If the data is potentially useful, hidden, trivial, these methods helps to find regularities. The aim of text categorization is to allow users to extract data from textual resource and to deal with operations such as retrieval, classification, clustering, data mining, natural language pre-processing and machine learning methods together to classify different pattern.

Text categorization can be stated like a set a= (a1 ,a2 ,...,ag ) where x is the jth document to be categorized and let set b=(b1 ,b2 ,...,bh ) where bk is the predefined category to which text document ak will be mapped for a function f. Here gdenotes the total number of documents that has to be categorized and the total number of pre-defined categories are denoted by h. It is represented like this [1]:

F: aj→ bk

With the rapid growth and development in technology, the research on text categorization has been evolved into a new stage, where the machine learning techniques have led to modality of text categorization. For example: K- Nearest Neighbor, Naïve Bayesian Classification, Support Vector Machine, Decision Trees etc.

## 2. ASSOCIATION RULE

Let S = {i1, i2, i3..., ik} be a set of attributes (items).
Let T= {t1, t2, t3..., tk} be a set of transactions (database).
Unique transaction ID to each transaction is assign which containing a subset of the item is S.

A rule is defined as below in an implication of the form:
P→Q
Where P, Q□S
And P∩Q=Θ.
Each rule contains two different set of items, which are also known as item sets, P and Q, where P is called antecedent and Q consequent.
Interesting rules are selected by using various rules, interests and constrains on various measures of significance. [2]

### 2.1 Support

The proportion of transactions in the database which contains the item set P is nothing but the support value of P with respect to T.
In formula: supp (P)

### 2.2Confidence

The confidence value of a rule P Q, with respect to a set of transactions T, is the proportion of the transactions that contains P which also contains Q.

Confidence is defined as:

Conf (P→Q) =support (P∩Q)/supp (P)

## 3. TEXT CLASSIFICATION

Given that most of the text in the modern era is in digital format, text classification plays an important role to manage and process such data. The goal of text categorization is the classification of documents into defined categories. The categories are nothing but just symbolic labels with no additional knowledge of their meanings. Fig 1 shows the different stages of text classification such as collection of documents, pre-processing, feature indexing, and feature filtering, different classification algorithm and performance measure.

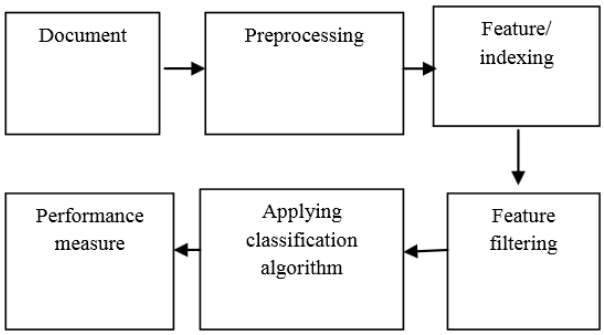


**Fig -1**: The Text Classification Process [3]

First step includes collection of data which can be present in various formats such as pdf, html, doc, php etc. Data which is collected goes through a series of steps: Removing stop words: Stop words are the common words that appear in text and carry little meaning, they serve only syntactic meaning but do not indicate subject matter; e.g. “the”, “a”, “and”, “that” are useless as indexing terms. A list of stop words is extracted and after scanning word by word, these words are removed.

Second step includes indexing of the document. Full text document is converted into document vector. It uses vector space model to represent documents which are nothing but the vectors of words.

In Feature-selection irrelevant features are removed for the purpose of classification and vector space is constructed to improve scalability, efficiency and accuracy of a text classifier.

Models describing important data are extracted using classification algorithms. Different techniques/methods such as K Nearest Neighbor, Artificial Neural Network, Naive Bayes Classifier, and Decision Trees are used to classify the models.

The last step for text classification is performance measure. Performance Evaluation of text classifier is done by calculating the precision and recall. Precision is the percentage of the documents that are retrieved that are in relevant to the query i.e. having a “correct” responses to the fired query. It is defined as

Precision =| {Relevant} {Retrieved}| {Retrieved}

Recall: this is the percentage of document that is relevant to the documents that are relevant to the query and were in fact, retrieved. It is formally defined as [7]

Recall = | {Relevant} {Retrieved}| {Relevant}

F measure is calculated by taking harmonic mean of precision and recall. [7]

F measure=Recall precision (Recall + Precision)/2.

## 4. CLASSIFICATION ALGORITHM

### 4.1 Naive Bayesian Classification

Naive Bayes is a simple and easy technique for classification: Naive Bayes classifier calculates posterior and prior probabilities to find the particular class. The Bayes’ Theorem is:

P (M|N) =P (N|M).P (N) P (M)

Where,

M- Some hypothesis, such that data tuple N belongs to specified class C

N – Some evidence, describe by measure on set of attributes

P (M |N) – the posterior probability that the hypothesis M holds given the evidence N

P (M) – prior probability of M, independent on N

P (N|M) – the posterior probability that of N conditioned on M.

Advantages*:*

- This technique work well on numeric as well as textual data.
- This classifier is easy to implement and computation are simple comparing with other algorithms.
- As it can be applied to large data set, no complicated iterative parameter estimation schemes are needed.
- Easy interpretation of knowledge representation.
- Performs well and it is robust.

Disadvantages*:*

- Does not consider frequency of word occurrences.
- Theoretically, naive Bayes classifier have minimum error rate when compared with other classifier, but practically it is not always true, because of the assumption of class conditional independence.

### 4.2 K-Nearest Neighbor

The k-nearest neighbors’ technique is based upon the principle that the samples which are similar to each other will lie in close proximity. Given an unlabeled sample, K-nearest neighbor classifier will search the pattern space for k-objects that are closest to it and will delegate the class by identifying the class label which is frequently used. For value of k=1 the samples which are closest to unknown samples are given pattern space. [4][5]

Advantages:

- Performs well on applications which has a sample with many class labels.
- This classifier is robust to noisy training data.
- Classifier is effective if the training data is not small.
- Vary little information is needed to make it work.
- Learning is simple.

Disadvantages:

- Slower than other classification examples.

- Nearest neighbor classifiers delegates equal weight to each attribute. This may cause confusion when there are many irrelevant attributes in the data and thus results into poor accuracy.
- It lacks to choose the value of N, except through cross-validation, which makes finding optimal value of N very difficult. [4]

### 4.3 Decision Tree

In decision tree "root" is the main nodewhich has no incoming edges from any other node. All the other nodes in the tree have exactly one incoming edge. Other nodes are called as internal nodes have only outgoing edges. Each nodes splits into two or more nodes.These nodes are called as terminal nodes.

Advantages:

- Decision tree has excellent speed of learning and speed of classification.
- Supports transparency of knowledge/classification.
- Supports multi-classification.

Disadvantages:

- Small amount of variations in the data can imply very different looking trees.
- Construction of a decision tree may affect badly for irrelevant attributes.

## 5. COMPARATIVE STUDY OF CLASSIFICATION ALGORITHMS

**Table -1:** Comparison of Classification Techniques

| | Naive Bayes | K - Nearest Neighbor | Decision Trees |
|---|---|---|---|
| Accuracy in general | Average | Good | Good |
| Speed of learning | Excellent | Excellent | V. Good |
| Speed of classification | Excellent | Average | Excellent |
| Tolerance to missing values | Excellent | Average | V. Good |
| Tolerance to irrelevant attributes | Good | Good | V. Good |
| Tolerance to noise | V. Good | Average | Good |
| Attempts for incremental learning | Excellent | Excellent | Good |
| Explanation ability/ transparency of knowledge/ classification | Excellent | Good | Excellent |
| Support Multi Classification | Naturally Extended | Excellent | Excellent |

Based on above comparison one can choose anyone of the above technique for text based categorization according to their need and performance requirement.

## 6. CONCLUSIONS

Text classification has a vital role in managing and processing data. In this paper, comparison amongst the classification algorithms like decision trees, Bayesian network and K-nearest neighbor has been done in detail. The aim behind this study was to find appropriate classification technique for text based categorization. The comparative study has shown that each algorithm has its own set of advantages and disadvantages and its own area of implementation. All the criteria cannot be fulfilled by a single classification algorithm. Two or more classifiers can be integrated and built according to the implementation and performance needed.

## ACKNOWLEDGEMENT

We acknowledge with gratitude to our Guide, professors, and our principal who has always been sincere and helpful in making us understand the varied concepts.
Apart from this, the term paper will be of immense importance for those who are interested in this subject. We hope they find it comprehensible.

## REFERENCES

[1]. S. Ramasundaram and S.P. Victor, "Algorithms for Text Categorization: A Comparative Study ", World Applied Sciences Journal 22 (9): 1232-1240, 2013.
[2]. Tan, Steinbach, Kumar, "Data Mining Association Analysis: Basic Concepts and Algorithms", 2004.
[3].The Wikipedia website. [Online]. https://en.wikipedia.org/wiki/Association_rule_learning [4]. Anuradha Patra, Divakar Singh," A Survey Report on Text

Classification with Different Term Weighting Methods and Comparison between Classification Algorithms", International Journal of Computer Applications (0975 – 8887) Volume 75–No.7, August 2013

[5]. Amit Ganatra, Hetal Bhavsar, "A Comparative Study of Training Algorithms for Supervised Machine Learning", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-4, September 2012.

[6]. Cover, T., Hart, "Nearest Neighbor Pattern Classification", IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, 1967.

[7]. Hetal Bhavsar, Amit Ganatra, "A Comparative Study of Training Algorithms for Supervised Machine Learning", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-4, September 2012

[8]. Anuradha Patra, Divakar Singh, "A Survey Report on Text Classification with Different Term Weighing Methods and Comparison between Classification Algorithms", International Journal of Computer Applications (0975 – 8887) Volume 75– No.7, August 2013.

## BIOGRAPHIES

**Omkar Vijay Ardhapure**
Student at ABMSP'S Anantrao Pawar College of Engineering & Research Pune.

**Gayatri Patil**
Student at ABMSP'S Anantrao Pawar College of Engineering & Research Pune.

**Disha Udani**
Student at ABMSP'S Anantrao Pawar College of Engineering & Research Pune.

**Kamlesh Jetha**
Assistant Professor at ABMSP'S Anantrao Pawar College of Engineering & Research Pune.