**Imperial Journal of Interdisciplinary Research (IJIR)**
Vol.2, Issue-1 , 2016
ISSN : 2452-1363 , www.onlinejournal.in

# A Survey on Text Classification of Documents Using Hybrid Techniques of Machine Learning.

**Nihar Ranjan**
*SITS, Narhe.*
nihar.pune@gmail.com

**Kavyashree Pushpan**
*SITS, Narhe.*
kavyasreepushpan09@gmail.com

**Shraddha Samgir**
*SITS, Narhe.*
shraddhasamgir728@gmail.com

**Anjali Nair**
*SITS, Narhe.*
nair.anjali294@gmail.com

**Rutuja Murhekar**
SITS, Narhe.
murhekarrutuja02@gmail.com

## Abstract

*Document classification is a problematic situation in the field of science. Text classification is just one task of text mining. Complete task of text mining is to give the user, the benefit of the textual information and user could be able to perform the task of text retrieval, classification and summarization. The problem for document classification is to assign a particular document to a specific category. This can be done manually or by using algorithms. Manually it may take more efforts to do this job since it is not possible to sort a huge number of documents manually. Thus, to reduce these efforts, a system is introduced that uses text classification algorithms. Text classification is basically classification of text document on the basis of some pre-defined genres and parameters. This paper introduces the algorithms like k-means, neural network, Decision tree and NLP.*

***Keywords****-Text mining, Text classification, ID3, Neural network, NLP.*

.

## 1. Introduction

Text classification has always been an important application. In last some years, text classification has gained a prominent status in the computer field. Today, text classification is a need due to the large amount of documents that is generated daily. Now-a-days the emerging numbers of documents are countless. All these documents are to be labelled under some category to make the searching of these documents easy. Thus text classification has a major role in today's life. As big data is a major problem in today's life so is the text classification, since all the big data produced are not necessarily formed under some specific category. These data are to be arranged properly. Otherwise just imagine the situations of search algorithms those use labels as their index, and search algorithms using labels as their index are very common. Thus text classification is an important task to be done. In general, text classification includes classification based on text and genre-based classification. Topic-based categorization categorizes the documents according to the topics of the input document. Texts can be found in many genres, for example: articles that relate to science, news reports, movie reviews, and commercials/advertisements. Genre is defined on many characteristics such as the way a text was created, the audience assigned to it, editing of that text as well as the author who created it and many more. We came to know that this task differs from topic-based categorization from the previous work in text classification. The common approach to build a text classifier is to manually label/assign a set of documents to pre-defined categories, and then use a machine learning algorithm to produce a classifier. This classifier can then assign classes to future documents based on the words/keywords they contain. The approach followed to build a text classifier is

commonly called supervised learning because the training documents have been labelled with pre-defined classes.

Text Classification is the task of classifying/assigning a document to a predefined category. To be precise, if d is a document of the entire set of documents D and {c1,c2,c3… } is the set of all the categories, then text classification assigns one category c to a document d. In detail it is all about detecting the genre of the unlabeled document. An unlabeled document is given as an input to the system. This document has no label for it and expects from the system to give a genre to this document. So the system scans that document and sends it to pre-processing unit. The document can be collected in various format .pdf, .html, .doc, etc. All these documents could be fetched from different sources. The pre-processing unit tokenizes, delete the stop words and produce a bag of words. Tokenizing means the entire document is scanned and divided into small number of tokens, so that these tokens can be collected individually. Delete the stop words means the words like "a", "the", "of", "an", "am", "and", "or", "is" and many such words are removed thus leaving a bag of useful words. System is already trained for categorizing the document using some keywords. The system has 5 pre-defined categories, for an example politics, movies, computers, etc. Each of these categories has decided their own keywords like for the genre or category movies, names of actors, box office are all keywords. Similarly each category has their specific keywords. With the help of these keywords, the system detects the genre of the unlabeled document. If the inputted document has the keyword of politics genre then the label of that document is decided as politics. After deciding the genre, the label of the document is set and hence the document is classified. The system also learns the newly added document and trains itself to pick out some more keywords, thus improving the efficiency of the system. For instance, if the genre of the document is decided upon as movies and the label is also set as movies then the system scans the document again and picks out some more new keywords on the basis of frequency count. The new document is thus scanned and new keywords are picked up. This improves the efficiency of the system. And also makes the system dynamic in nature.

## 2. Literature Survey

Now a day's, text classification has become a topic of interest to every individual. In the following we examine some basic and advance review papers related to text classification.

In this paper [1] they have discussed about the basic architecture of the text classification process. The process includes reading the document, tokenizing the text, stemming, deleting the stop words, vector representation of text, feature selection and feature transformation and learning algorithms. Each of this term is explained in a very well manner. All the details are provided for each term. In feature transformation there is a method named as Latent Semantic Indexing (LSI), which is used for indexing. A combination of KNN and LSI are also applied for feature transformation. The next thing this paper recited about was of learning algorithms. The machine learning algorithms specified in this paper are naïve bayes, rule induction, neural networks, decision trees, nearest neighbours, support vector machine. Decision trees, naïve bayes, nearest neighbours are all the oldest techniques. As per the paper, Naive Bayes is most commonly used technique in text classification applications and experiments because of its simplicity and effectiveness. But its performance is degraded or we can say its performance is very less since it cannot handle text well. The paper says that Support vector machines (SVM), when applied to text classification provide excellent or best precision, but poor or less recall. The paper also included some evaluation part. In this part they tell us how to determine effectiveness; however, precision, recall, and accuracy are mostly used. To determine these, one must first start by understanding if the classification of a document was a true positive (TP), false positive (FP), true negative (TN), or false negative (FN). These terms were fully explained by the paper. In this paper [2], they have described the document classification processing which start as following. First Document Collection, followed by pre-processing, indexing, feature selection, classification algorithms, performance measure. Here they have explained the various classification and clustering algorithms. Classification algorithms are Sequential decision tree based classification, Parallel formulation of decision tree based classification, neural network, and genetic algorithm. Clustering algorithms are hierarchical method and partitioning method. They have also explained about algorithms for discovering associations. These all were types of text mining algorithms. More such text mining algorithms are concept mining, information retrieval, information extraction. Thus this paper described the different text mining algorithms in detail. In this paper [3] they discussed about the text classification and classifiers. They also explained the document classification process which ended with performance

evaluation. In this last step of text classification the efficiency of classifier is calculated. As described in the paper, for finding estimates of precision and recall relative to the whole category set, two different methods may be used like Micro-averaging and Macro-averaging, some other measures are also used as Break–even point, F-measure, Interpolation. Many classifiers are explained in this paper which are Rocchio's Algorithm, K-Nearest Neighbours, Naïve Bayes, decision tree, decision rule, neural network, LLSF, voting, Associative classifier, Centroid based classifier. All these classifiers were explained thoroughly. All algorithms are good for classification; even hybrid classifier can also be used. A comparative study is taken place in the paper. In this paper [4] text categorization techniques are explained. Also the applications of text categorization are specified. Applications are Spam Filtering, Automatic indexing, Document Organization, Text Filtering, Word Sense Disambiguation, and Hierarchical Web Page Categorization. The types of text categorization discussed were Single-label vs. multi-label text categorization, Category-pivoted vs. document-pivoted text categorization, and Soft versus Hard Text Categorization. The categorization methods specified are decision tree, Bayesian, n-gram, vector-based. Different feature selection methods are also cited such as Document Frequency, Information Gain, Mutual Information, and Chi Square. Thus this paper gave a lot of information related to text categorization applications, types and methods. In this paper [5], they completely described the support vector machine method of text classification. In this classification method, use of vector machine is given. In this paper [6], the base of text mining that is information retrieval is explained thoroughly. Information retrieval is a process of extracting information from different sources and providing them to user to do text analysis and classification. By retrieving information analysing of text becomes easy and also classification becomes easy.

In this paper [7], the detailed work of the system that classifies the text is given. The general work flow of the system is as follows. First create a matrix, create container, training models, classify the data using training models, analytics and lastly testing the accuracy of the system. In this paper [8] they have discussed about the partially supervised learning methods. They have cited some of the theorems for partially supervised learning. Here, they have also mentioned a term of positive documents and negative documents. Basically this paper aims to find the class of a document from a group of mixed documents. Thus the term positive document implies the document which is needed or required, and the term negative document implies the rest of the documents from the mixed set of documents. In this paper [9], they have mentioned that by combining one or two classifiers, the classifier accuracy could be increased. Thus, they have cited three various combination approaches, namely, simple voting, Dynamic Classifier Section (DCS) and their own approach of adaptive classifier combination (ACC). Simple voting is a common method, where the number of classifiers individually assigns the test document to the specific class. In DCS, k-nearest neighbour and 'leave-one-out' method is combined. ACC is similar to DCS but instead of choosing the best classifier with the highest local accuracy, they choose the class that has highest local accuracy.

## 3. Text Classification Method.

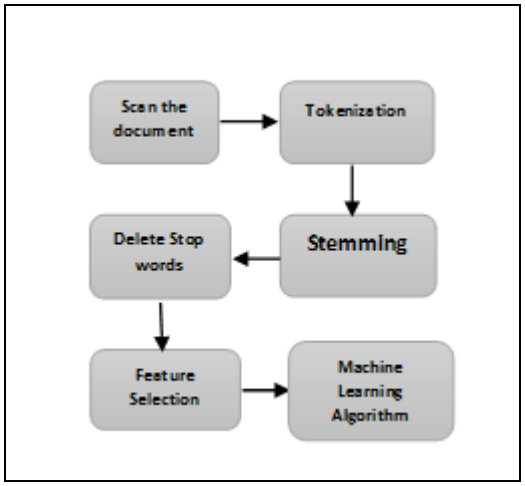


**Figure 1. Basic Architecture**

### 3.1. Document Scanning

This is the primary step of classification process, where the document, like .pdf, .doc, .html, is taken as an input and it is scanned well and forwarded to the pre-processing unit.

### 3.2. Tokenization

In this step of classification, the whole document is tokenized. This means the document is divided in to numbers of small tokens.

### 3.3. Stemming

This step does the job of removing the misspelled words or a word with the same stem, i.e., the original word is only counted. Like the words connecting, connection, connectionless has the same stem "connect". Thus only the word "connect" is kept in the document and rest of the words are deleted. Here the word "connect" is the root of every other word like connecting or connection.

### 3.4. Delete Stop Words

Here the useless words like "the", "an", "a", "and", etc. are removed or deleted from the document. Many such prepositions are deleted.

### 3.5. Feature Selection

After pre-processing and indexing the important step of text classification, is feature selection. It is used to construct vector space, which improves the scalability, efficiency and accuracy of a text classifier.

### 3.6. Machine Learning Algorithm.

**3.6.1. Neural Network.** A neural network is an algorithm where input set is a number of terms while output set contains the genre or category. For classification of a test document, terms weights are assigned to the input set, which is propagated forward along the network by which output weights are decided which gives the conclusion of the category of the test document. Perceptron is used to map the input weight to the output weight; it basically maps the input to the network that leads to the specific output. Single-layer perceptron and multi-layer perceptron are the two available options. But multi-layer perceptron is used mostly.

Advantages:

- Simple to implement.
- A neural network can perform those tasks that a linear program cannot perform.
- A neural network has a good technique of network learning and it does not need reprogramming.
- These are more efficient.
- Neural network can be executed in any application and no problem is faced.

Disadvantages:

- Neural network are hard to be retrained. If you add data later, then it is difficult to add to an existing network.
- Handling time series data is a tough job in neural networks and sometimes it is impossible.
- To operate, a neural network needs to be trained.
- Neural network's architecture is different from the architecture of microprocessors thus it needs to be emulated.
- Requires high processing time for large neural networks.

**3.6.2. ID3.** ID3 is a decision tree algorithm which stands for "Iterative Dichotomiser 3". A decision tree classifies a problem by diving them from the root node to some leaf node. Each attribute defines a test conducted on the problem and their following branches indicate the possible solutions. ID3 uses top-down, greedy search to testify each of the attribute.

Advantages:

- ID3 implicitly perform feature selection or variable screening.
- For data preparation, decision trees require comparatively less effort from users.
- They are easy to interpret.
- It represents step by step solution, which can be explained easily.
- Programming language is not a constraint.

Disadvantages:

- It is very time consuming.
- If the data is classified incorrectly, then it cannot be updated, instead a new tree is to be generated.
- Numeric data cannot be handled.
- Only one attribute at a time is tested.
- Logic expression cannot be understood by ID3.

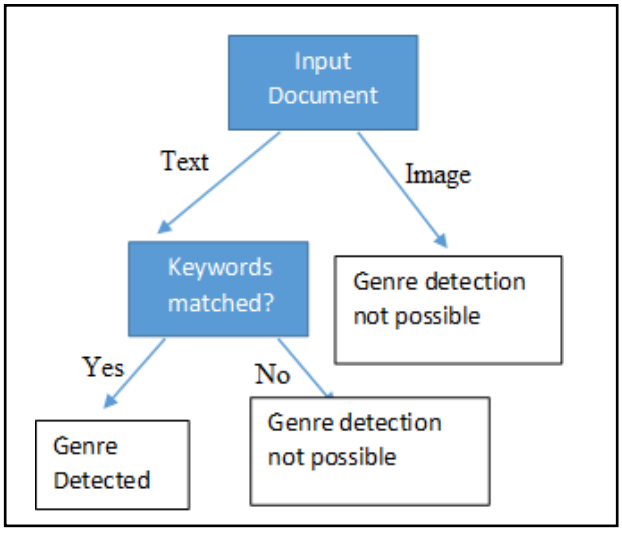


**Figure 2. Decision Tree**

**3.6.3. NLP.** Natural language processing (NLP) is a field of, artificial intelligence, which deals with the interactions between computers and human (natural) languages.

**3.6.4. K-means.** Clustering problem is found in many different applications, such applications are data mining, knowledge discovery and data compression. Clustering based on k-means is related to the location problem. It simply estimated the mean of the set of k-groups. Initially k-object are chosen and are named as cluster seed, these represent the temporary mean of the cluster. Then square Euclidean distance from each cluster seed to the each object is calculated and after that each object is assigned to the closest cluster. For each newly created cluster, a new centre is calculated. Each seed value is replaced by the respective cluster centre. Again the mean distance from each object to the centre is calculated and according to that new clusters are formed. Repeat above process until the mean remain the same.

## 4. Conclusion

In this paper we have done a survey on neural network, decision tree, k-means and NLP. These are some of the machine learning algorithms that is used for text classification. We observed that by combining these algorithms, text classification process could be more easy and efficient. We also studied about the basic text classification process. The document is scanned and tokenized. From these tokenized words, stop words are deleted, feature selection is done and a bag of words are produced. These words are classified using any of the classifier mentioned above.

## 5. References

[1] M. Ikonomakis, S. Kotsiantis and V. Tampakas, "Text Classification Using Machine Learning Techniques", WSEAS TRANSACTIONS on COMPUTERS, Issue 8, Volume 4, August 2005, pp. 966-974.

[2] 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 (IJAIEM) Volume 2, Issue 3, March 2013.

[3] Vandana Korde and C Namrata Mahender, "TEXT CLASSIFICATION AND CLASSIFIERS: A SURVEY", International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012.

[4] Meenakshi, Swati Singla, "Review Paper on Text Categorization Techniques", SSRG International Journal of Computer Science and Engineering (SSRG-IJCSE) – EFES April 2015.

[5] Thorsten Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features", Dortmund, 27.November, 1997 Revised: 19. April, 1998.

[6] R. Thamarai Selvi, E. George Dharma Prakash Raj, "An Approach to Improve Precision and Recall for Ad-hoc Information Retrieval using SBIR Algorithm", 2014 World Congress on Computing and Communication Technologies.

[7] Timothy P. Jurka, Loren Collingwood, Amber E. Boydstun, Emiliano Grossman, and Wouter van Atteveldt, "RTextTools: A Supervised Learning Package for Text Classificatio", The R Journal Vol. 5/1, June ISSN 2073-4859.

[8] Bing Liu, Wee Sun Lee, Philip S. Yu, Xiaoli Li, "Partially Supervised Classification of Text Documents".

[9] Y. H. Li and A. K. Jain, "Classification of Text Documents", THE COMPUTER JOURNAL, Vol. 41, No. 8, 1998.