# NFCKE: New Framework for Document Classification and Knowledge Extraction

**Ghanshyam Singh Thakur**[1] and **Dr. R. C. Jain**[2]

***Abstract - In this research paper we have developed NFCKE text categorization systems. Text classification has many applications, such as fraud detection, automatic email classification and web-page categorization. The results show that NFCKE does better than other widely used techniques. The new framework for document classification is efficient and accurate. Our experiments indicate that the accuracy of existing method increase by using this approach. The final goal is achieving high performance and eventually increasing classification accuracy.***

***Index Terms - Text mining, classification, Binary Matrix Model (BMM).***

## 1. INTRODUCTION

Document classification has been studied intensively because of its wide applicability in areas such as web mining, information retrieval. The majority of this information is in text format, for example, emails, news, web pages, reports, etc. Organizing them into a logical structure is a challenging task. More recently, classification is employed for browsing a collection of documents or organizing the query results. Although standard classification techniques such as k-means, Support Vector Machines, Naive Bayes, Decision Trees[15,16,18,19], can be applied to document classification, they usually do not satisfy the special requirements for classification documents: all these approaches are suffer from lack of high performance and high accuracy. In addition, many existing document classification algorithms require the user to specify the number of category as an input parameter. Incorrect estimation of the value always leads to poor classification accuracy. Furthermore, many classification algorithms are not robust enough to handle different types of document sets in a real-world environment. In some document sets, category sizes may vary from few to thousands of documents. This variation tremendously reduces the resulting classification accuracy for some of the state-of-the art algorithms. But there are still problems to be tackled such as efficiency and accuracy. Owing to wide significant applicability of text categorization and challenges in the area motivated us to do work in this field. The poor classification accuracy and the weaknesses of the standard classification methods formulate the goal of this research. We provide an accurate, efficient, and scalable classification method that addresses the special challenges of document classification. NFCKE is a relatively new concept comparatively more efficient and accurate. So for, no research was conducted to use NFCKE concept for Document classification. This approach seems promising because we can apply different method of classification. BMM is the out come of NFCKE. We applied existing classification methods on the BMM. Our experiments indicate that the accuracy of existing method increase by using the model (BMM). We can also apply associative classifier on BMM Model, which generate association rules between keywords or features. The final goal is achieving high performance and eventually increasing classification accuracy. All these issues motivated and directed our research.

[1, 2] *Department of Computer Applications, Samrat Ashok Technological Institute, Vidisha(M.P.), INDIA*

## 2. METHODOLOGY

The main goal of this research is to develop high performance and new optimization text categorization algorithms that will reduce the time complexity and space complexity of algorithms and finding various applications of the text categorization algorithm in real world problem as the result of computerization. The amount of text documents available in digital form has been growing significantly during the last decades due to the development of new technology. These include e-mail, newsgroups and on-line news, all of which can be stored in text form. The accelerating growth in the amount of text data makes it necessary to automate. In this paper we limit our attention to document classification accuracy and high performance.

Text classification is based on supervised learning model. In this learning we divided our dataset into two parts. One part is called training dataset and another part is called test dataset. With training dataset we create a model or classifier. Once we created a classifier we estimate the accuracy of the classifier using test dataset. For mining large document collections it is necessary to pre-process the text documents and store the information in a data structure, which is more appropriate for further processing than a plain text file. The aim of this paper is to preprocess documents, to apply classification method and improve accuracy of text categorization using NFCKE, which is the basis for various applications such as e-mail classification and Web-page classification. NFCKE is a new concept for Text categorization witch includes several sub-phases that should be integrated for efficient and accurate outcomes. These sub phases include document collection, preprocessing, indexing, feature selection, model preparation and estimation of classifier accuracy and performance. We explore these sub-phases in terms of an approach to similarity based text categorization. After performing sub phases like preprocessing, indexing and feature selection of NFCKE, we apply BMM-based text classification method because, which is fast and more robust method compared to others text classification methods. BMM-based text classification is one of the new supervised approaches to classify texts into a set of predefined classes with relatively low computation. New framework documents are almost unstructured and all classifications algorithm requires structured form of the

documents. So it is the mandatory to convert unstructured document into a structured document representation. In our new framework we explain in details this conversion. This new framework includes following sub-phases-

1. Document Collection-we collect relevant documents for classification.
2. Preprocessing-in this phase we perform the operation like removal of HTML Tags, Special character, stop words and perform word stemming
3. Indexing- in this phase we perform the operation of assigning the weight to feature
4. Feature Selection- in this phase we perform reduce dimension of documents
5. BMM Representation- to represent classes into two dimension tabular form.
6. Performance and accuracy: we estimate the performance and accuracy of BMM method

After performing sub phases like preprocessing, indexing and feature selection of NFCKE, we apply BMM-based text classification method because, which is fast and more robust method compared to others text classification methods. BMM-based text classification is one of the new supervised approaches to classify texts into a set of predefined classes with relatively low computation

## 3. DOCUMENT REPRESENTATION

All the algorithms applied in Text Classification need the documents to be represented in a way suitable for the inducing of the classifier. For the document classification we represent using Binary Matrix Model (BMM) as binary matrix. In Binary Matrix each rows represent a documents and each columns represent term in the documents. Number of rows indicates number of document and number of column indicate number of terms in the document.

### 3.1 Binary Matrix Model (BMM)-

This model represent by a matrix called binary matrix.Binary Matrix M is represented as

- M[di × wj] = 1 , if wj Є di
- = 0, otherwise
- Where i=1,2,…..n
- j=1,2,3,…..m

Binary Matrix Model (BMM) is a powerful approach. This model is based on the binary values, 0 represents the absence of the term in the document and 1 represents presence of the term in the document. This model can used wide variety document classification algorithm.

## 4. EXPERIMENTAL RESULT ANALYSIS

We have done the experiment on 20 Newsgroups data sets—This dataset is a collection of 20,000 newsgroup documents, partitioned into 20 different newsgroups. The 20 different news groups are as follows -
alt.atheism,comp.graphics,comp.os.ms-windows.misc, talk.politics.misc, comp.sys.mac.hardware,comp.windows.x,misc.forsale,rec.autos,rec.motorcycles,rec.sport.baseball,rec.sport.hockey,sci.crypt,sci.electronics,sci.med,sci.space,soc.religion.christian,talk.politics.guns,talk.politics.mideast,talk.religion.misc,comp.sys.ibm.pc.hardware

| S.No. | BMM | k-nn | Bayes |
|---|---|---|---|
| Datase1 | 91.81 | 83.98 | 84.98 |
| Datase2 | 92.16 | 84.14 | 86.14 |
| Datase3 | 91.17 | 84.76 | 85.76 |

**Table 1**

To improve the classification performance, we adopt the following method each data set was split into 80% and 20% for training and testing respectively. A 3-fold cross validation was carried out to determine the training set accuracy at the various parameter settings. The parameter that gave the best training accuracies was then used to determine the accuracy values for each of the corresponding test sets. To compare the performance of the classification methods, we look at a set of standard performance measures. .

**Figure 1**

In this learning we divided our dataset into two parts. One part is called training dataset and another part is called test dataset. With training dataset we create a model or classifier. Once we created a classifier we estimate the accuracy of the classifier using test dataset. BMM-based classifier consistently and substantially outperforms other algorithms such as Naive Bayesian, k-nearest-neighbors, and C4.5, on a wide range of datasets. BMM -based text classification presents accuracy close to that of the state-of-the-art methods.

## 5. CONCLUSION

In this paper we have developed a new framework for classification and a new BMM classifier. We conducted extensive comparative experiments on standard test collections (the 20-Newsgroups). We experimentally predict that a BMM model which is the outcome of NFCKE give high accuracy and efficiency for classification. We also experimentally showed that BMM gives high accuracy and performance than other classification methods. The results show that NFCKE does better than other widely used techniques.

***Continued on page no. 59***