ISSN No. 0976-5697

Volume 6, No. 2, March-April 2015

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Document Classification using Neural Networks Based on Words

Chaitanya Naik
Computer Engineering
K. J. Somaiya College of Engineering (KJSCE)
Mumbai, Maharashtra, India

Vallari Kothari
Computer Engineering
K. J. Somaiya College of Engineering (KJSCE)
Mumbai, Maharashtra, India.

Zankhana Rana
Computer Engineering
K. J. Somaiya College of Engineering(KJSCE)
Mumbai, Maharashtra, India

***Abstract*:** Categorization is the process of classifying the documents into various predefined categories called as classes. A category is chosen considering the relation between the subject of the category and the document belonging to it. Document categorization may include classification of text, images, audio etc. There is huge information being stored in various electronic forms and hence, a proper classification of documents is necessary to achieve an organized data. This paper explains classification of documents into predefined classes using **neural networks** with the use of MATLAB tool.

***Keywords*:** Document classification, neural networks, training, testing.

I. INTRODUCTION

In the last few years, the use of digital documents for storing and accessing the information has increased to a very great extent[1]. Storing data in electronic form has number of advantages like availability of space for storage as well easy availability of the document anywhere and at anytime. However, with this easiness of the use of electronic form, an important issue regarding this vast amount of information is to organize the information and easy retrieval of huge amount of data. Document classification is the key technique in text mining to organize the information in a supervised manner. Document classification is a task of classifying the documents into predefined categories [1]. Digital documents may be in the form of text, audio, video. In this paper, we focus on classification of a digital text document stored in .txt, .doc, .docx format. Various algorithms are used for this classification. Document classification may be done using either Rule-based or Machine-learning approach. We primarily focus on the use of neural network technique to classify the document [1].

Neural network is a machine learning approach to classify the documents. In the proposed work, the keywords are extracted from the document and these are used for the classification purpose.

II. LITERATURE REVIEW

The task of document classification may be achieved using two approaches[2]. One is Rule-based approach and the other one is the Machine-learning approach. Rule-based approach is the one in which the documents are classified manually using the if-then rules. The advantage of this approach is that it has high precision but the disadvantage of this approach is poor recall and poor flexibility. It becomes a tedious task to classify huge amount of data manually and hence, is a time-consuming approach. Due to the above disadvantages of the Rule-based approach, we do not discuss this approach in detail and focus only on the Machine-learning approach[2].

The methods used in Machine-learning approach for the classification problem are K-Nearest Neighbour (KNN), Support Vector Machines (SVM), Naïve Bayesian (NB) and Neural Networks (NN). Out of the above four methods, KNN is the simplest method for classification. KNN[2] is a classification algorithm in which the objects are classified into classes considering the smallest distance between the class and the object. However, the disadvantage of KNN is that it costs very much time for classifying objects if number of training examples is large because it has to select few objects by computing the distance of each test object using all the training examples. The second approach to text categorization is NB [2]. It is different from KNN by the fact that it is trained using the training examples in advance to classify the unseen examples. It classifies documents based by calculating prior probabilities of the predefined categories/classes and probabilities that attribute values belong to categories. Here, we assume that attributes are independent of each other and hence, underlies on this approach. However, this assumption violates the fact that attributes are dependent on each other in a text classification application. The next approach is the use of SVM for text categorization [2]. This is a more popular machine-learning algorithm than the other two mentioned above. SVM is based on the idea of linear classifier, perceptron which is an early neural network. The idea of SVM is different from that of perceptron model in the sense where if a distribution of training examples is not linearly separable, then these examples are mapped into another space where their distribution becomes linearly separable [2].

III. NEURAL NETWORK

A. What is a neural network [3]?

A neural network is a processing device, either an algorithm or an actual hardware whose design was inspired by the design and functioning of animal brains. The power of human brain comes from the sheer number of neurons and their multiple interconnections. It also comes from genetic programming and learning. There are over 100 classes of neurons. The individual neurons are complicated and convey information via a host of electrochemical pathways. Together those neurons and their connection form a process which is not binary, not stable and not synchronous. In short it is nothing like currently available electronic computers or even the artificial neural network itself. An Artificial Neural Network (ANN) is information-processing model inspired by the biological nervous system, such as thebrain to process information. This model replicates only the most basic functions of the brain. ANNs[3] possess large number of highly interconnected elements called nodes or neurons. They usually operate in parallel and are configured in a regular architecture (Fig. 1). ANNs collective behavior is recognized by their ability to learn, recall and generalize the training patterns or data similar to that of a human brain.

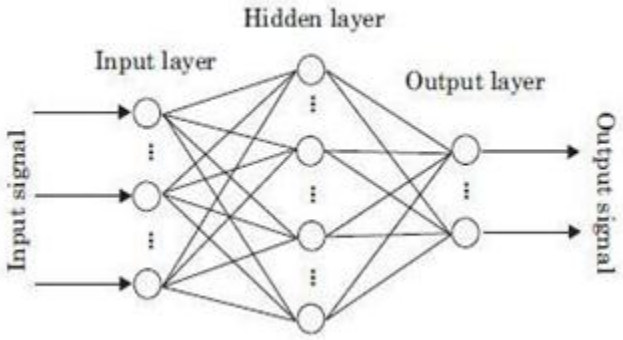

Fig. 1 : Architecture of neural network

B. Why ANN?

We are trying to find the solution to the Text categorization problem by Backpropagation method, which is a technique of Artificial Neural Network System (ANS)[4]. There is a strong reason for using ANS in text categorization. For the problems which cannot be solved sequentially or by sequential algorithms ANS provides the better solution. Apart from Text categorization, for the other applications like pattern matching of images, non sequential algorithms are provided by artificial neural network. These algorithms are better in performance. Among the various algorithms provided by ANS[4], Backpropagation is a very popular algorithm. Backpropagation as an ANS is very useful in recognizing complex patterns and performing nontrivial mapping functions. Following figure represents a simple Backpropagation diagram. The rounded objects represent the neurons or processing elements of a neural network. The directed lines that are connecting the neurons are called weights. Also every line of processing elements is a layer of a network. Thus in this figure there are three layers available. Though there can be multiple layers present in ANS generally there will be three layers present in Back-Propagation network (BPN) as shown in fig 2.

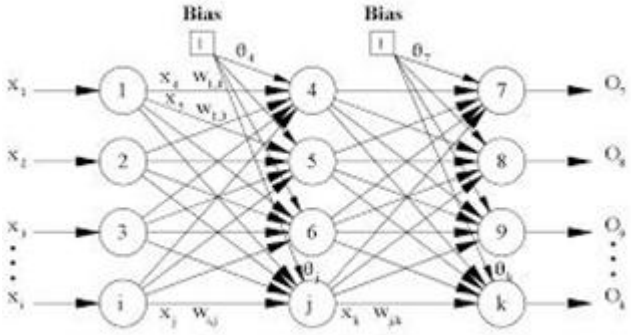

Fig. 2 : Back-propagation network

C. Back-Propagation algortihm.

One of the most important developments in the neural networks is Back-Propagation learning algorithm. The network which is associated with this algorithm is called Back-Propagation network (BPN). This network is applied to the multilayer (input, output and hidden) feed-forwarding networks which consists of processing elements with differentiable continuous activation function. For a given set of training input-output pair, a procedure is provided by this algoithm for changing the weights in a BPN to classify the given input patterns correctly. The basic concept used for this weight update algorithm is simply the gradient-descent method. This is a method where error is propagated back to the hidden unit. The aim of the neural network is to train the net to achieve a balance between the net's ability to respond and the ability to give reasonable responses to the input that is similar and not identical to the one that is used for training.

The BPN algorithm is produced in two phases here. In the first phase forward signal propagation occurs in the network. In the second phase the error terms are fed back to all other input units. In this case they are the feature vectors. Now the algorithm provided below[5]:

1) **Phase 1: Propagation:** Each propagation involves the following steps:
- Forward propagation of a training pattern's input through the neural network in order to generate the propagation's output activations.
- Backward propagation of the propagation's output activations through the neural network using the training pattern target in order to generate the deltas of all output and hidden neurons[5].

2) **Phase 2: Weight update:** For each weight-synapse follow the following steps:
- Multiply its output delta and input activation to get the gradient of the weight.
- Subtract a ratio (percentage) of the gradient from the weight[5].

There are certain aspects worth mentioning in BPN. The first thing is that BPN is good at generalization. Here generalization means BPN will learn to eliminate significant similarities in the input vectors if the different input feature vectors belonging to a same class are given. Irrelevant data

will be ignored. The second thing is that if the output function is sigmoidal, then we have to scale the output values. Because of the sigmoid function, the network outputs can never reach 0 or 1. Therefore use values such as 0.1 and 0.9 to represent the smallest and largest output values.

IV. CLASSIFICATION ALGORITHM

The algorithm we have implemented for document classification uses neural network for the training and the testing purpose. The classification of the document includes a sequence of steps to be performed. 1. Preprocessing (Removal of Stopwords and Stemming). 2. Find the keywords from the unclassified document. 3. Apply the classification algorithm.

A. Preprocessing

Preprocessing of the document includes two steps viz. Stopwords removal and Stemming.

Stopwords Removal[6]: Stop words do not have so much meaning in a retrieval system and are a part of natural language. Stop-words should be removed from a text because texts look heavier and less important for analysis. Removal of stop words results in reduction of the dimensionality of term space. Prepositions, articles, and pro-nouns etc are the most common words are in text documents that does not provide the meaning of the documents. They are known as stop words. Some of its examples are: the, in, a, an, with, etc. The reason for elimination of Stop words from the document is they are not considered as keywords in text mining applications[6].

Stemming: For finding out the root/stem of a word, stemming technique is used. Stemming results in conversion of words to their stems which incorporates a great deal of language-dependent linguistic knowledge. For example, the words, driving, driven, drive, drove all can be stemmed to the word 'drive'. In the present work, the Porter Stemmer algorithm is used which is the most commonly used algorithm in English[6].

Apart from the above two steps, we need to preserve the paragraphs in a document and hence, we introduced a delimeter between two paragraphs in a document in this step itself. The reason for introduction of delimeter between the paragraphs will be explained in the next section. The delimeter we used in our experiment was ### (three #s) at the end of each paragraph. However, for further research or experiments, you can use any special symbol, letter or a combination of letters and/or symbols as delimeter.

B. Find the keywords

Keywords are the words extracted from an unclassified document based on which the document will be classified into one of the predefined categories. The selected keywords are passed through a neural network and we get the category to which that word is classified at the output of the neural network. Now, here comes the role of the delimeter. The technique that we are using for document classification is a branch and bound method where the document is first divided into paragraphs; paragraphs are preprocessed and each word in a paragraph is considered individually. However, to classify a document based on words, we need to choose the important words in a document called keywords. So, at first every keyword chosen from a paragraph will be classified in a category and based on those words, a paragraph will be classified; and finally based on categories to which every paragraph is classified, the entire document is classified. We have chosen three keywords from every paragraph; viz. 1. Highest frequency word present in the considered paragraph. 2. Lowest frequency present in the considered paragraph. 3. Average frequency [(highest+lowest)/2] word present in the considered paragraph. In case of a clash between two or more words for highest, lowest or average frequency, we had chose the word on FCFS basis i.e. to choose that word as a keyword which first appeared in the paragraph. An alternative for this can be to choose the word which is alphabetically first. Now the three keywords chosen from a paragraph are given as input to the neural network.

C. Role of neural network.

A neural network is trained in order to later classify the documents. The training phase of the neural network is done using the Backpropagation algorithm as explained earlier.

Training of the neural network: The training of the neural network is the most important part of the task of classification using neural network. For the training of the neural network, a training dataset is to be provided to the neural network. A training dataset contains the features based on which any object is to be classified. In our experiment, we had to classify the documents into three categories viz. Physics, Chemistry and Biology (The domain for our experimentation purpose was science related documents); so for the documents to be classified into one of these three categories, we provided a vocabulary list for each category as the training set. We created a database which contained a vocabulary list for each of the three categories i.e. the database had a table with three columns Physics, Chemistry and Biology; we added words to each of the columns considering their relevance to that subject. For example: words like ‘physicist’, ’electromagnetism’, ’radiation’ etc were added to the column Physics; words like ‘reaction’, ‘chemical’, ‘thermodynamics’ etc were added to the column chemistry and lastly words like ‘gynaecology’, ‘life’, ‘cells’ etc were added to the column Biology. Now using this vocabulary the neural network was trained. Since we require numerical values to give as input to the neural network, each letter in the word was converted to its ASCII code and then given to the neural network. For example: physicist was converted to and stored as ‘p’ ‘h’ ‘y’ ‘s’ ‘i’ ‘c’ ‘i’ ‘s’ ‘t’ : [112 104 121 115 105 99 105 115 116]. Now for the neural network, we chose three layers; 1 input layer (26 neurons), 1 hidden layer (26 neurons) and 1 output neuron. Since we had set fixed number of input neurons and the input to this layer was the vocabulary list of varying size, we normalized every word to length 26 by appending n number of zeros after the word ended; where ‘n’ is normalized length(26 in our experiment) minus the length of the word. Now the training set is given to the neural network. A target

has to be set during the training phase of the neural network. In our experiment, we added 1000 words of each category and were given to the network in the sequence physics followed by chemistry followed by biology; so the target value was set as 1 for first 1000 words, 2 for the next 1000 words, 3 for the next 1000; which means target value for physics is 1, for chemistry is 2 and for biology the target value was set to 3. (The above target values can be changed as per the requirements or the number of categories.) And then the neural network was trained. For the implementation of neural network we used the tool MATLAB; and the training of the network was carried out using the Backpropagation algorithm. Once the network is trained, it is now ready to classify any unclassified document.

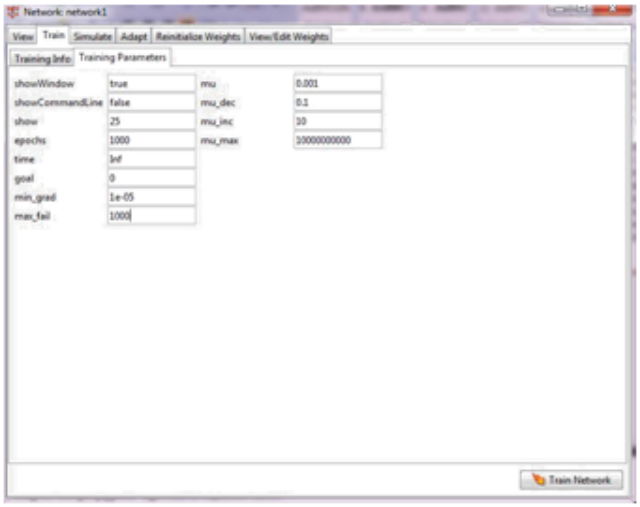

Fig. 3: Training parameters

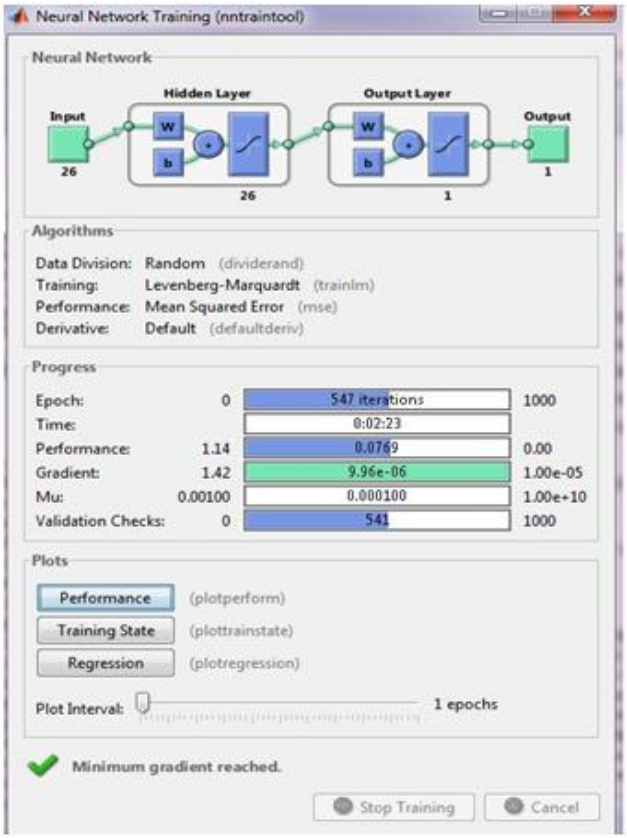

Fig. 4: Training of the neural network

Fig. 4 shows the training of the network.

Stoping Training Criteria[7]:

During training, the progress is shown in the training window. The criteria considered are the performance, the magnitude of the gradient of performance and the number of validation checks. To terminate the training, the magnitude of the gradient and the number of validation checks are performed. As and when the training reaches a minimum of the performance, the gradient will become very small[7]. The training will stop as magnitude of the gradient becomes less than 1e-5. There can be an adjustment made by setting the parameter net.trainParam.min_grad. The number of validation checks is the number of successive iterations that the validation performance fails to decrease. The training will stop when this number reaches 6 (the default value),. In this run, you can see that the training did stop because the gradient limit had reached[7]. (The results may differ than those shown in the following figure, because of the random setting of the initial weights and biases.)

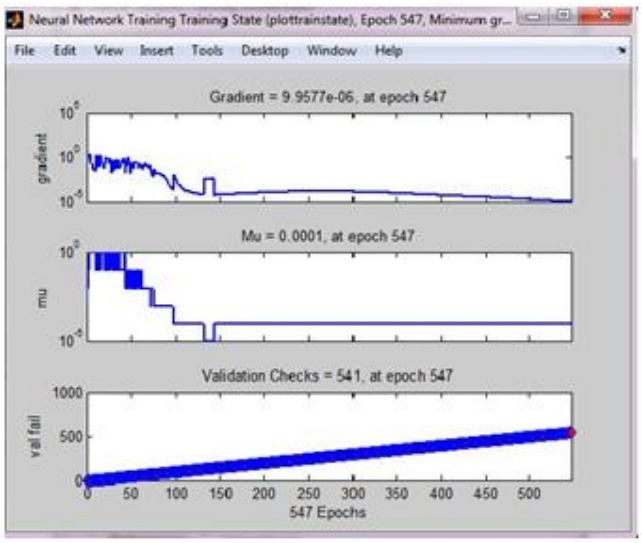

Fig. 5: Neural network training graph

Fig. 5 shows the graphical representation of the training phase of the neural network.

Testing of the neural network and classification: Now, the trained network can be used for classification purpose. The keywords obtained from the second step are to be given to the neural network. The words are converted to their corresponding ASCII codes and are normalized to length 26 as defined in our experiment. The three keywords are given to the neural network one by one and the output for each is recorded. The output of each word will be close to 1 or 2 or 3. These outputs will be used to classify the paragraph as described in table below:

Table 1: Output

Physics	Chemistry	Biology	Classification
0	0	0	Not classified
0	0	1	Biology
0	1	0	Chemistry
1	0	0	Physics

Table 1 shows that if out of the three more than one word belong to a single category, the paragraph is classified in

that category. The classification of the paragraph is stored and the same procedure is repeated for the remaining paragraphs of the document. Later, when all the paragraphs are classified, the category to which the majority paragraphs belong will be the category to which the document will be classified. However, if there is a word which does not appear in the training set, then for the following word no result will be obtained close to the target values. Then in such a case the paragraph cannot be classified. The word is stored in w and that paragraph at this moment is not classified and the remaining paragraphs will go through the same above classification steps. Based on the outputs of the remaining paragraph, we classify the document to the category to which majority paragraphs excluding the one which has the word not present in our training set belong. Thus, the document is classified. Now, the word initially not present is to be added to the vocabulary list and hence that word is to be trained by the neural network. Since we now know where the document is classified, we get to know to which category the word w belongs to. So we choose the corresponding target value for word w and the word is trained. Thus, the vocabulary list is updated by a new word to one of the categories.

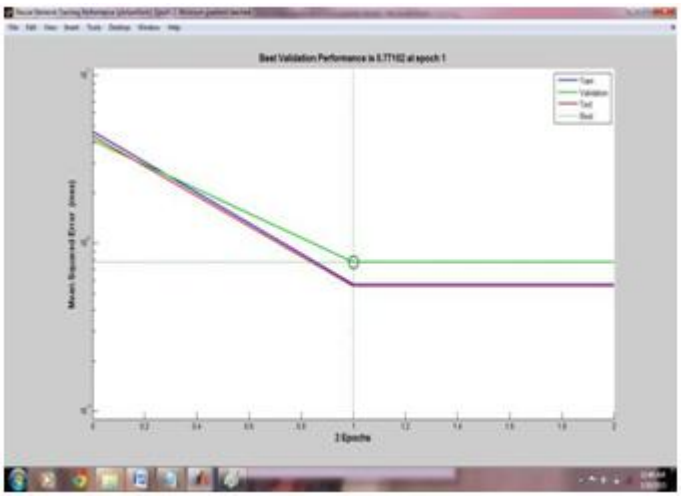

Fig. 6: Performance graph

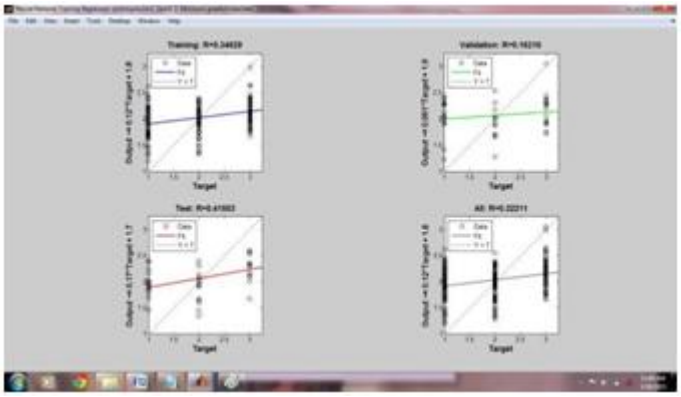

Fig. 7: Regression graph

Fig. 6 is the performance graph of the trained network and it plots the training, validation, and test performances given the training record TR returned by the function train. Fig. 7 is the regression plot and it plots the linear regression of targets relative to outputs.

Hence, document classification is achieved as a supervised learning process where the documents are classified based on a predefined vocabulary list as well the neural network is self learning as it updates the list by training those words which initially did not appear in the list; and once updated this word can now be used to classify any other document in which this word appears.

V. APPLICATIONS

The document categorization is mainly used for accessing the wanted document in a sophisticated manner so that in future the data or the document itself can be modified and retrieved preserving all its semantics and attributes[8]. In the business world, document categorization is used so that the data can be stored in data repository such that it is secured. In industrial field, it is used for betterment of information storage[8]. Using this document categorization technique, the binarization process can be implemented to extract data from corrupted documents (specially the ancient manuscript).

VI. CONCLUSION AND FUTURE WORK

The above algorithm is a simple and efficient algorithm in order to classify the document. The domain we chose for our experiment was restricted to documents related to science. However using the same algorithm, a hierarchical structure of classification can be obtained like a science document is first classified into chemistry; further a chemistry document can be classified to organic or inorganic chemistry; organic can be further classified to thermodynamics or fluid mechanics and so on. Apart from science as the domain, documents relating any other subject or category can be classified by passing a vocabulary list corresponding to the required category through the neural network and get trained and then use it in classification of other related documents. Also the variation in choosing of the keywords can be used to obtain better results. This classification algorithm works well with .txt, .doc, .docx files and also word converted to pdf files. A scanned pdf document is first subject to image processing and then the same document classification algorithm can be applied to it.

VII.ACKNOWLEDGMENT

Our sincere thanks to the expert Mrs Rajani Pamnani, who have contributed towards the working of this experiment.

VIII. REFERENCES

[1]. Menaka S, Radha N, "Text Classification using Keyword Extraction Technique", International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, December 2013 ,Issue 12, ISSN: 2277 128X

[2]. Taeho Jo, "NTC (Neural Text Categorizer): Neural Network for text categorization", International Journal of Information Studies Volume 2 April 2010 Issue 2.

[3]. S.N. SIVANANDAM, S.N. DEEPA, "Principles of Soft Computing, second edition."

[4]. S.Ramasundaram, S.P.Victor, "Text Categorization by Backpropagation Network", International Journal of Computer Applications (0975 – 8887) Volume 8– No.6, October 2010.

[5]. http://www.wikipedia.com

[6]. RUWAN GAMAGE, "An Ontology Based Fully Automatic Document Classification System Using an

Existing Semi-Automatic System", IFLA WLIC 2013, SINGAPORE.

[7]. Amit Ganatra, Y P Kosta, Gaurang Panchal, Chintan Gajjar, "Initial Classification Through Back Propagation In a Neural Network Following Optimization Through GA to Evaluate the Fitness of an Algorithm", International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011.

[8]. Debnath Bhattacharyya, Poulami Das, Debashis Ganguly, Kheyali Mitra, Purnendu Das, Samir Kumar Bandyopadhyay, Tai-hoon Kim, "Unstructured Document Categorization: A Study" International Journal of Signal Processing, Image \Processing and Pattern Recognition.