A Novel Semi Supervised Algorithm for Text Classification Using BPNN by Active Search

Mahak Motwani[1],Aruna Tiwari[2]

**[1] Assistant Professor, Computer Science Department, TCST
Bhopal,M.P.,462032 , India**

**[2] Assistant Professor, Computer Science Department, IIT
Indore, M.P, India**

Abstract

Demand of Text Classification is increasing with the evolution of huge amount of text data available in internet, news, institutes , To make an effective text classifier we need large amount of labeled data in the form of training samples, to get labeled data is not only expensive but also time consuming, tedious task, whereas unlabelled data is easily available & inexpensive. This paper proposes an algorithm that just makes use of some root words from expert followed by active search. Our algorithm also makes use of a very effective Term weighting method based on relevance factor that is used for feature representation, this text is train by BPNN. The proposed algorithm is compared on test data and on standard data 20 Newsgroup and mini Newsgroup on the basis of micro-average and macro-averaged F_1 measure The Experimental results depicts the best micro averaged F_1 measure of 0.95 at 2400 epochs for test data, 0.67 for 20 news group and is 0.95 for Mini Newsgroup which are comparable with the well known supervised Text classification.

.

Keywords: *Semi Supervised, text classification, Active search, term weighting method, Neural network*.

1. Introduction

With the immense growth of online text data .Text Classification is one of the most essential requirements for effective navigating, summarizing and organizing data. Text classification is the process of classifying text in one or more category. There are many algorithms proposed for Automatic Text classification.

Depending on the data available text classification can be categorized as supervised or unsupervised. Supervised learning is learning from labeled examples. It is an area of machine learning that has reached substantial maturity, it has generated general purpose and practically successful algorithms[1], whereas learning without the use of samples or labeled data is unsupervised learning where one finds an interesting structure with sample independently drawn from unknown distribution, Unsupervised learning is closely related to the problem of density estimation in statistics[2], major issues of unsupervised learning are minimum domain knowledge, noisy data, insensitive to instance order etc

The need for large quantities of data to obtain high accuracy, and the difficulty of obtaining labeled data led the Research direct towards field where one can use a lot of unlabeled data which are easily available rather than labeled data which are manually assigned by experienced analyst which makes it time consuming and labour intensive job.

Semi supervised is a way to make use of this huge amount of easily available unlabeled data and few labeled data which makes it perform better than unsupervised algorithm, our proposed algorithm is making use of only few root words and easily available relevant data to train the classifier.

Our proposed algorithm apply active search makes use of just few root words and gets web assisted labeled data this Text data is pre-processed and filtered that reduces the number of dimension and a fine term weighting method is applied to represent the text data as per the relevance between terms and document, this data is used to train the classifier based on the standard BPNN , Section 2 briefly describes about Pre-processing Term weighting method, about active search and neural network classifier , section 3 discusses the proposed algorithm, section 4 introduces the experimental methodology, section 5 reports experimental results and discussion and section 6 concludes the paper with future work

2. Proposed Framework

2.1 Pre-processing & Term Weighting Method:

Major issues related to text data are dealing with unstructured text, handling large number of attributes, examining pre-processing techniques for text classification, semantic and natural language processing based techniques and choice of a suitable machine learning technique for training a text classifier. Text documents has huge number of dimension since we consider each word as a dimension handling such text is extremely complicated, since the text document are semi structured, the first and vital step is effective text representation which converts the content of text document into compact format so that the document can be recognized by a classifier.

Data pre-processing reduces the size of the input text documents significantly. It involves activities, natural language specific stop-word elimination [3] [4] [5] and stemming [4] [6]. Stop-words are functional words which occur frequently in the language of the text (for example, "a‟, ‟the‟, ‟an‟, ‟of‟ etc. in English language), they are not useful for classification. Stemming is the action of reducing words to their root or base form. For English language, the Porters stemmer is a popular algorithm [6] [7], which is a suffix stripping sequence of systematic steps for stemming an English word, reducing the vocabulary of the training text by approximately one-third of its original size [6].For example, using the Porters stemmer, the English word "globalizations" would subsequently be stemmed as "globalizations → globalization → globalize → global". Our algorithm uses porter's stemming algorithm for suffix stripping

Feature extraction / selection helps identify important words in a text document, the paper proposes feature extraction of the text by filtering according to the term frequency(tf), i.e. frequency of each term in document (occurrence of the term in document)is evaluated and then only those words which occur in more than one document are considered and thus reduces the feature .

Now each word in a document contributes to the semantic of the category differently and has different importance in text. The term weighting methods assigns an appropriate weight to the term to improve the performance of text classification[8] paper investigates several widely used unsupervised and supervised term weighting methods, a new simple supervised term weighting method, tf,rf, (term frequency, relevance frequency)is used to improve the terms' discriminating power for text categorization task, here emphasis has been made on term discriminating power analysis ,relevance factor refers to the degree of relevance of the term to the category it belongs to as compared with its relevance to other documents. It has been proved that it has a consistently better performance than other term weighting methods while other supervised term weighting methods based on information theory or statistical metric perform [19].

In text classification of multiple classes, a term may have high term frequency (t.f) and may belong to almost all the classes in this case the term actually do not posses a high discriminating power and so the inverse term document frequency factor and its variant has been used, our proposed algorithm uses a supervised term weighting method which is a multiplication of t.f and relevance factor r.f. where relevance factor is defined as

$$r.f= \log (2+ (a/\max (1, c)) \qquad (1)$$

Here

a: total number of document in the positive category that contain this term

c: number of document in the negative category that contain this term

Here we assign a term as positive category if it belongs to the document that belongs to the category and all other categories combined together as negative category

2.2 Categorization by active search

Labelling large amount of text spans for training systems is time consuming and unrealistic for many applications. We consider here the use of semi-supervised techniques, which allow to train a system with only a few labeled documents together with large amounts of unlabeled documents, It is difficult to build reliable classifier that is able to achieve high classification accuracy with of small number of available labeled documents, one way to overcome this problem is by using active search.

Active search is a way to first identify a number of important keywords, root words belonging to different category and then utilize search engines to retrieve from the web a multitude of relevant documents [9], we use Google to get relevant documents.

Though initially we have unrelated keywords, query word the web data or document collected will undergo effective Pre processing and feature selection term weighting method to remove the irrelevant words and proceed for training.

2.3 Learning

Supervised learning, unsupervised learning and semi supervised

In supervised learning also called directed data mining, the variables under investigation can be split into two groups: explanatory variables and one or more dependent variables. The target of the analysis is to specify a relationship between the explanatory variables and the dependent variable. To apply directed data mining techniques the values of the dependent variable must be known for a sufficiently large part of the data set. The training data contains examples of inputs together with the corresponding outputs, and the network learns to infer the relationship between the two. Training data is usually taken from historical records, this data in which the class, or the category the data belongs to is already known is labeled data

Unsupervised learning is a class of problems in machine learning where the goal is to determine how data is organized. Many methods employed here are based on data mining methods used to pre-process data. It is distinguished from supervised learning in that the learner is given only unlabeled examples.

Semi-supervised learning is a learning paradigm concerned with the study of how computers and natural systems such as humans learn in the presence of both labeled and unlabeled data. Traditionally, learning has been studied either in the unsupervised paradigm (e.g., clustering, outlier detection) where all the data are unlabeled or in the supervised paradigm (e.g. classification, regression) where all the data are labeled. The goal of semi-supervised learning is to understand how combining labeled and unlabeled data may change the learning behaviour, and design algorithms that take advantage of such a combination. Semi-supervised learning is of great interest in machine learning and data mining because it can use readily available unlabeled data to improve supervised learning tasks when the labeled data are scarce or expensive. Various classification techniques have been studied and analysed in the literature based on fuzzy [10] Semi-supervised learning algorithms that using large amount of unlabeled data, together with the labeled data, to build better classifiers is an open and interesting problem. Such approaches include EM algorithm [11][13][14], Co-training [11][12][13],Co-EMT algorithm[13][15].

2.4 Classifier Using Neural Network

There exist a number of approaches to text categorization machine learning based, heuristic and rule based approaches. Since heuristic approaches are based on rules of thumb, they are more accurate and precise but have poor recall, tolerance and require a time consuming job of building rules manually, they are not very appropriate for real data which is full of noisy data.

Typical machine learning algorithms applied traditionally to text categorization are KNN (K Nearest neighbour), NB (Naïve Bayes), SVM (Support Vector Machine), and BP (Back Propagation). The four approaches to text categorization have been used more popularly in previous literatures on text categorization than any other traditional approaches.

Our algorithm uses neural network for text classification since it supports massive parallelism and its good when interacting with noisy data, its fault tolerance and can adapt with change in circumstance, since we are dealing with text data which is highly unstructured and has huge dimension, an algorithmic solution cannot be defined for it, using neural network for text data is an appropriate solution.

Learning in neural is a process by which free parameters of neural network are adapted through a process of stimulation by the environment in which the network is embedded. The type of learning is determined by the manner in which the parameter changes take place" [16] A number of neural network model can be used for supervised learning, Among the various algorithms provided by Neural network, Back propagation is a very popular algorithm. Back propagation as an ANS is very useful in recognizing complex patterns and performing nontrivial mapping functions, Though there can be multiple layers present in ANS generally there will be three layers present in Back propagation network

The neural network is a three-layer fully connected feed-forward network which consists of an input layer, a hidden layer and an output layer. All neurons in the neural network are non-linear units with sigmoid function as the activation function. In the input layer, the number of input units (r) is equal to the dimensionality of the reduced feature space. In the output layer, the number of output units (m) is equal to the number of pre-defined categories in the particular text categorization task. The number of hidden units in the neural network affects the generalization performance. The choice depends on the size of the training set and the complexity of the classification task the network is trying to learn, and can be found empirically based on the categorization performance.

3. Proposed Algorithm

Our Algorithm starts with some keywords, root words belonging to the category of computer science and sports, root words from expert, and then using active search, retrieve the relevant documents using search engine [9] , algorithm starts with 34 root words of computer , 42 root words of sports category and 39 words of Medical using these words retrieve100 documents belonging to computer science, medical and sports category.

These document undergo the process of pre processing, text tokenization process is called as word extraction, word breaking, word segmentation, or lexical analysis. Depending on the natural language the document is written in, the word extraction process involve different techniques. In English language it is very easy because boundaries between words are marked by special delimiting characters such as spaces and punctuation. A set of functional words called stop words (is, am ,a ,the …etc) occur frequently but do not posses any importance since these are non content words that appear in all the document, such words are removed , the words and its various forms like verb, adjective their tense are reduced to their roots, this process of stripping suffix is Stemming, Porters stemmer is used in proposed algorithm. Filtering is the process of reducing dimension by removing less relevant terms, terms which appear only in one document are removed.

Feature Representation: Set of features and their impact describe the properties of the objects have a greater impact on the classification of the objects. Hence to improve the accuracy of the classifier, identifying a set of "good" features representation sometimes take lots of time and becomes a critical step in constructing the classification system. a term may have high term frequency (t.f) and may belong to almost all the classes in this case the term actually do not posses a high discriminating power and so the inverse term document frequency factor and its variant has been used, our proposed algorithm uses a supervised term weighting method which is a multiplication of t.f and relevance factor r.f. where relevance factor is defined as

r.f= log (2+ (a/max (1, c)) ;where a is the number of times the term appear in documents of positive category and c is the number of documents the term appeared in negative category.

Back Propagation Neural Network

Multilayer feed forward network which uses a supervised learning method, a generalization of delta rule is known as back propagation learning algorithm .Back propagation neural network (BPNN) is the most popular in all of the neural network applications. It has the advantages of yielding high classification accuracy. The training of a network by back propagation involves three stages: the feed-forward of the input training pattern, the calculation and back-propagation of the associated error, and the adjustment of the weight and the biases.

Input pattern feed-forward. Calculate the neuron's input and output. For the neuron j, the input Ij and output Oj are

$I_j=\sum W_{ij} * O_{j;}$ (2)

$O_{j=}f(I_j+ \theta_j)$ (3)

where wij is the weight of the connection from the ith neuron in the previous layer to the neuron j, $f(I_j+ \theta_j)$is an activation function of the neurons, Oj is the output of neuron j, and θ_j is the bias input to the neuron. In this paper, we use a tanh(n)sigmoid activation function defined with the equation:

tansig(n) = 2/(1+exp(-2*n))-1; (4)

This function is a good trade off for neural networks. The error, E, is calculated in this paper, the mean absolute error function is used in the output layer The mean absolute error is used to evaluate the learning effects and the training will continue until the mean absolute error falls below some threshold or tolerance level.

$$E = \frac{1}{2\pi} \sum_n \sum_l \sqrt{(T_{nl} - O_{nl})^2} q \quad (5)$$

Here n is the number of training patterns, l is the number of output nodes, and O_{nl} and T_{nl} are the output value and target value ,respectively. The mean absolute error is used to evaluate the learning effects and the training will continue until the mean absolute error falls below some threshold or tolerance level. The back propagation errors both in the output layer, δ_l and the hidden layer, δ_j, are then calculated with the following formulas:

$$\delta_l = \lambda(T_l - O_l) f'(O_l)$$
$$\delta_j = \lambda \sum_i \delta_l W_{ij} f'(O_j)$$

(6)

Here T_l is the desired output of the l_{th} output neuron, O_l is the actual output in the output layer, O_j is the actual output value in the hidden layer, and k is the adjustable variable in the activation function. The back propagation error is used to update the weights and biases in both the output and hidden layers.

Weights and biases adjustment: The weights, wji, and biases, θi, are then adjusted using the following formulas:

$$W_{ji}(K+1) = W_{ji}(k) + \eta\delta_j O_i \quad (7)$$

$$\theta_i(k+1) = \theta_i(k) + \eta\delta_i \quad (8)$$

Here k is the number of the epoch and g is the learning rate.

The back propagation error is used to update the weights and biases in both the output and hidden layers[17]

4. Experimental Methodology

34 Query words of Computer science field,39 Query word of Medicine and 40 Query words of Sports are used to retrieve 100 documents of the three fields that are divided in 70 to 30 ratio of training and test documents, This training documents undergo Tokenization i.e removal of special character, numeric values etc, after tokenization we get 363985 terms from 210 training text files, followed by removal of 428 stop words, porter stemming algorithm is applied the terms reduces to 16768 words are retrieved that are filtered on the basis of occurrence in number of document, we retrieve only those terms that occur in more than one document and thus 6458 terms are considered.

Term weighting method based on relevance factor is used for feature representation, this data belonging to three category computer science ,medicine and sports undergoes training in BPNN with following parameters The parameters used for BPNN are neurons in input layer and 20 neurons in hidden layer, training function used is gradient descent adaptive training function, tansig as activation function for hidden layer and linear function for output layer, learning rate used is 0.3, momentum of 0.6.

4.1 Evaluation Criteria

Precision and Recall are two popular performance measures for text classification, precision is the fraction of retrieved documents that are relevant, recall is the fraction of relevant documents that are retrieved. The set of documents that are both relevant and retrieved is denoted as relevant ∩ retrieved,

Precision= Relevant ∩ retrieved/retrieved (9)

Precision = true positives / (true positives + false positives)(10)

Recall: This is the percentage of document that are relevant to the documents that are relevant to the query and were in fact, retrieved. it is formally defined as

Recall= relevant ∩ retrieved/relevant (11)

Recall =true positives/(true positives+ false negatives) (12)

However, neither precision nor recall makes sense in isolation from each other as it is well known from the IR practice that higher levels of precision may be obtained at the price of low values of recall. To combine precision and recall, the two most widely used measures, i.e micro-averaged F_1 and macro-averaged F_1 Score.

These are two conventional methods of calculating the performance of a text categorization system based on precision and recall. The first is called *micro-averaging*, while the second one *macro-averaging*. Micro-averaged values are calculated by constructing a global contingency table and then calculating precision and recall using these sums. In contrast macro-averaged scores are calculated by first calculating precision and recall for each category and then taking the average of these. The notable difference between these to calculations is that micro-averaging gives equal weight to every document it is called a document-pivoted measure. while macro-averaging gives equal weight to every category ,category-pivoted measure[18]. micro-averaged F_1 and macro-averaged F_1 Score is the Harmonic mean of corresponding Precision and Recall.

$$P_{\text{micro}} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} TP_i + FP_i}; \quad R_{\text{micro}} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} TP_i + FN_i} \quad (13)$$

$$P_{\text{macro}} = \frac{1}{|C|}\sum_{i=1}^{|C|} \frac{TP_i}{TP_i + FP_i}; \quad R_{\text{macro}} = \frac{1}{|C|}\sum_{i=1}^{|C|} \frac{TP_i}{TP_i + FN_i} \quad (14)$$

4.2 Experimental results and discussion

Initially In this Study, experiments under two experimental circumstances are performed. The performance of proposed algorithm on web assisted test data belonging to three category(30 test files/each category) computers, medicine and sports retrieved based on term frequency(tf), term frequency .relevance frequency(tf.rf) and term frequency . inverse document frequency (tf.idf) on the basis of Micro-Averaged F_1 measure Figure 1 amd Macro-Averaged F_1measure Figure 2 on different epochs is as shown

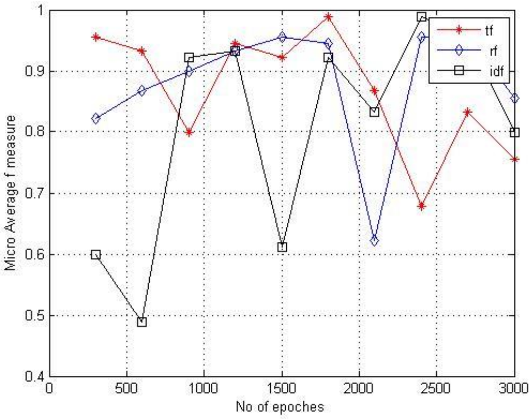

Fig. 1 Results on the Test data set

Experiments on the first series is shown in Fig 1 Micro-averaged F_1 measure on term weighting method based on term frequency and term frequency*relevance frequency, Term frequency. Inverse Document frequency.

Analysis on the basis of different no of epochs depicts that with increase in the training of neural network, Micro Averaged F_1 Measure increases .On an average it shows that relevance factor perform better than term frequency, and inverse document frequency also the results for both the measure is quite comparable. The trends of curve of micro averaged and macro averaged are almost similar for the methods.

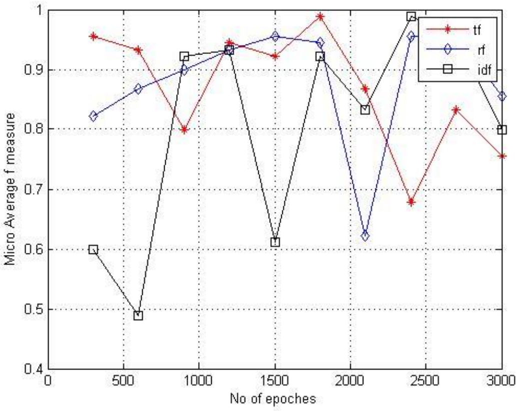

Fig. 2 Results on the Test data set

Test on random set of 10 to 50 document files of each category of computer and sports of standard 20 newsgroup test data The 20 Newsgroups corpus3 is a collection of approximate 20,000 newsgroup documents nearly evenly divided among 20 discussion groups and mini newsgroup data set is also standard data for text Mining. Data of category Computer science and Sports is analysed and micro averaged F_1 measure corresponding to different sets of input is shown in figure 3 and figure 4.

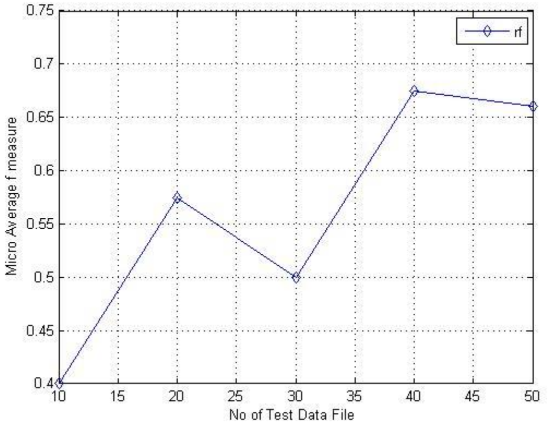

Fig. 3

The Presented algorithm performs well, the best micro averaged F_1 points 0.6750 for 20 news group and is 0.95 in Mini Newsgroup ,the algorithm gives better results for mini newsgroup data set.

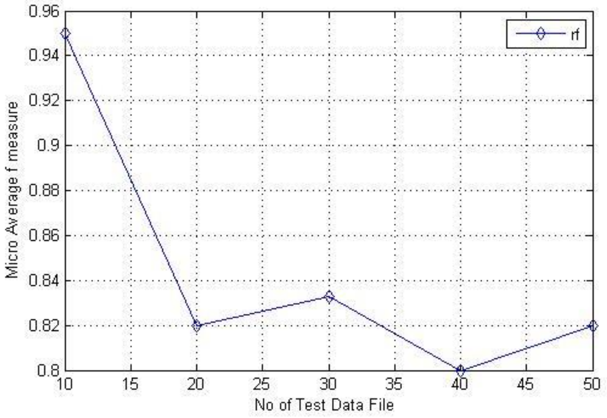

Fig. 4

6. Conclusion and Prospects

Our algorithm follows the step of pre processing, followed by an relevance factor, term weighting method and makes use of only few root words to train the classifier using active search rather than collecting huge amount of labeled data which is error prone, tedious and expensive to get, the implemented method performance give the micro averaged F_1 measure in the range of .8 to .9 for test data. Presented algorithm performs well with the standard data 20 newsgroup and mini newsgroup .Back propagation algorithm can be modified to perform better, research on innovative concept based representation can be used for text document representation

Acknowledgments

Specially thankful to my guide Dr. Aruna Tiwari, Assistant Professor at Indian Institute Technology, Indore who motivated, supported and guided me to complete this Research paper. Hearty thanks to worthy institute TCST, Bhopal and all those that we have inadvertently forgotten for their immense support and encouragement

References

[1] Maria-Florina Balcan, Avrim Blum,2010”A discriminative model of semi supervised learning” ACM DOI 10.1145/1706591.1706599

[2] Jordan, Michael I.; Bishop, Christopher M. (2004). "Neural Networks". In Allen B. Tucker. Computer Science Handbook, Second Edition (Section VII: Intelligent Systems). Boca Raton, FL: Chapman & Hall/CRC Press LLC.

[3] Kim S., Han K., Rim H., and Myaeng S. H. 2006. Some effective techniques for naïve bayes text classification. IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 11, pp. 1457-1466.

[4] Zhang W., Yoshida T., and Tang X. 2007. Text classification using multi-word features. In proceedings of the IEEE international conference on Systems, Man and Cybernetics, pp. 3519 – 3524.

[5] Hao Lili., and Hao Lizhu. 2008. Automatic identification of stopwords in Chinese text classification. In proceedings of the IEEE international conference on Computer Science and Software Engineering, pp. 718 – 722.

[6] Porter M. F. 1980. An algorithm for suffix stripping. Program, 14 (3), pp. 130-137.

[7] Goyal R. D. 2007. Knowledge based neural network for text classification. In proceedings of the IEEE international conference on Granular Computing, pp. 542 – 547.

[8] Lan M,Tan C L,Su J,Lu Y, Supervised and traditional term weighting methods for automatic text categorization, IEEE Trans Pattern Anal Mach Intell. 2009 Apr;31(4):721-35.

[9] Zenglin Xu, Rong Jin , Kaizhu Huang† Michael R. Lyu, Irwin King, Semi-supervised Text Categorization by Active Search, CIKM’08, October 26–30, 2008, Napa Valley, California, USA, ACM 978-1-59593-991-3/08/10.

[10] Girish Keswani and Lawrence 0. Hall “Text Classification with Enhanced Semi-supervised Fuzzy Clustering” Fuzzy Systems, 2002. FUZZ-IEEE'02. Proceedings of the 2002 IEEE International Conference .

[11] Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. Text classification from labeled and unlabeled documents using EM. Machine Learning, 39, 103–134. 2000.

[12] Blum A., Mitchell T., Combining labeled and unlabeled data with co-training. In: Proceedings of the Workshop on Computational Learning Theory, 1998.

[13] Nigam, K., & Ghani, R. Analyzing the effectiveness and applicability of co-training. Ninth International Conference on Information and Knowledge Management (pp. 86–93), 2000.

[14] Lu Mingyu, Research on Improvement of Bayes Text Classifier,Computer Engineering vol 32(17),(pp.63-65),2006,

[15] Muslea, I., Minton, S., & Knoblock, C. Active + semi-supervised learning = robust multi-view learning. Proceedings of ICML-02, 19th International Conference on Machine Learning (pp. 435–442) 2002.

[16] Simon Haykin,"Neural Network, A comprehensive Foundation",Pearson Education

[17] Combination of modified BPNN algorithms and an efficient feature selection method for text categorization Cheng Hua Li , Soon Cheol Park, Information Processing and Management 45 (2009) 329–340

[18] Vincent Van Asch "Macro- and micro-averaged evaluation measures " September 9, 2013

[19] Mahak Motwani,Aruna Tiwari" Comparative Study and Analysis of Supervised and Unsupervised Term Weighting Methods on Text Classification" International Journal of Computer Applications (0975 – 8887) Volume 68– No.10, April 2013

Mahak Motwani received B.E Degree in Computer science & Engineering from Ravi Shankar University, M.Tech in Computer science & engineering from RGPV, Bhopal. She is currently Pursuing PhD in the field of Data Mining from RGPV, Bhopal. She has been working from 2008 to 2013 as Assistant Professor in Computer Science Department of Truba institute of Engineering & Information Technology, Bhopal. Currently she is working as Assistant Professor in Computer science department of Truba College of Science and Technology, Bhopal, India

Aruna Tiwari received her B.E..and M.E. degree in computer science from SGSITS, Indore . PhD degree in Computer Science from RGPV, Bhopal. She worked as Lecturer in Shri Vaishanav Instt. Of Tech. & Sc., Indore from 1997 to 2001, she was working with SGSITS, Indore from 2001 to 2012 as Associate Professor, Currently She is working in Computer science department of Indian Institute of Technology, Indore, India.