International Journal of Scientific Research & Engineering Trends
Volume 4, Issue 3, May-June-2018, ISSN (Online): 2395-566X

An Artificial approach to Classify Text Document by Utilizing Clustered Terms

M. Tech. Scholar Shivangi Pandey
Dept. of Computer Science & Engineering,
Technocrats Institute of Technology, Bhopal India
Shivangi.2000pandey@gmail.com

Prof. Deepak Tomar
Dept. of Computer Science & Engineering,
Technocrats Institute of Technology, Bhopal India

***Abstract* – As the web clients are expanding, so advanced information on servers are increasing, this draws researcher from content mining field to upgrade different services. As different issues are emerge on the server, for example, information dealing, security, support, and so on. In this paper article characterization is recommended that group the document in proficient way. Here Error back propagation artificial neural network was used for the grouping. Proposed approach characterizes the information on the premise of text document features where preparing of neural network is perform by binary contribution of the content. While in testing stage content archives are order according to neural network output. Investigation is perform on genuine and in addition artificial dataset. Result demonstrates that Proposed work is better when compared with past work on various assessment parameters.**

Keywords*: *Classification analysis, Supervised Classification, Un-supervised Classification, Text Feature, Text Mining, Text Ontology,

I. INTRODUCTION

With the increase of digital text data on the servers. Text mining importance is increasing as this decrease lot of labor work for different use of text data. In this text mining research field classification of information and retrieval of documentation is highly required. So combination of various data mining techniques is done while gathering information from the text document [1]. As various researchers are working for improving accuracy of the work, but there is lot of improvement in the work for further increasing the parameters.

As text data is highly unorganized because it contain natural language. So mining for retrieval of information from text data is crucial for the researcher. Different pre, post, processing steps are taken for improving the information quality. While in case of text document information retrieval, it is found that most of the document data is open for all. Due to this privacy of the text dataset is very low. So this work has focus on two issue first is text information retrieval and second is privacy maintenance of the dataset. Ways to mine the text and cluster the documents for better processing is our concern.

Even any small activity of human produces electronic data. For example, when any person buys a ticket online, his details are stored in the database. As most of electronic or digital data available on servers are in text form. This data is highly un-clustered or structure less but also suffered from the large amount of waste information. In this data good quality of information is also available for the scientific and industry purpose. As most of the historic data is available in text which need to be update but this required skilled labor or reader how have knowledge of the different terms for conversion. So considering all these facts in 1960 Pittsburg University has requirement of computer enabled system is desired which perform these task efficiently. In mid 1960 university has develop a computer enabled research assistant for performing the text reading [5]. In this computer programs Boolean logics were set with nearness expression in form of phrase were used.

Text mining and data mining are similar, except data mining works on structured data while text mining works on semi-structured and unstructured data [9]. Data mining is responsible for extraction of implicit, unknown and potential data and text mining is responsible for explicitly stated data in the given text [2]. On the other hand potential information extraction is common to both [2].

In this paper text document mining algorithm is proposed which is a combination of neural network based clustering algorithms and other data mining techniques. Here terms are classified first then documents have been clustered into the most appropriate clusters, under which they belong most appropriately. The use of such text document mining techniques can be applied in dataset management, to maintain data quality. This work can be applied in order to cluster and classify the large number of documents that is unorganized. This is mainly required for easy access to the accurate document in minimum time. Thus, improving the mining techniques that can be used in the ever growing size of documents collected.

II. RELATED WORK

In this section few research work of this field is explained which specify different approaches of the researchers. Here it is found that classification of text document required good set of features for getting effective output.

© 2018 IJSRET

Dr. B. Poorna, Sudha Ramkumar in [1] has done text document clustering where grouping for a set of documents was done based on the information it contains and to provide retrieval results when a user browses the internet. In this work results shows that proposed work has retrieve the text document efficiently by prior classification of the text files in the document. Here work has focus on reducing the dimension of the dataset. So dimension reduction is done in by two approaches first is reducing of noise or text which do not provide any information while second is removing of unwanted features from the document dataset.

K. Fragos et al. in [2] also concludes in favor of combining different approaches for text classification. The methods that authors have combined belong to same paradigm – probabilistic. Naïve Bayes and Maximum entropy classifiers are chosen to test on the applications where the individual performance is good. The merging operators are used above the individual results. Maximum and Harmonic mean operators have been used and the performance of combination is better than the individual classifiers.

S. Keretna et al. [3] have worked on recognizing named entities from a medical dataset containing informal and unstructured text. For this, they combine the individual results of Conditional Random Field (CRF) classifiers and Maximum Entropy (ME) classifiers on the medical text; each classifier trained using a different set of features. CRF concentrates on the contextual features and ME concentrates on the linguistic features of each word. The combined results were better than the individual results of both the classifiers based on Recall rate performance measure.

S. Ramasundaram et al. [4] aimed to improve the N-grams classification algorithm by applying Simulated Annealing (SA) search technique to the classifier. The hybrid classifier NGramsSA brought about an improvisation to the original NGrams classifier while inheriting all the advantages of Ngrams approach. Feature reduction using method is used but its multivariate value among the n-grams affects the performance of the classifier.

III. PROPOSED METHODOLOGY

As the mining is utilized in different types of data analysis. So all need to increases the different technique in the required area. So proposed work contribute the text mining is by clustering the document or articles in the group without having any prior knowledge of the documents. In the proposed work no need of any format for the input data such as speaker's identification symbol or special character, here all process is perform by utilizing the different combination of terms and pattern features.

3.1 Preprocessing

Preprocessing is a process used for conversion of document into feature vector. Just like text categorizations the preprocessing also has controversy about its division [1, 7].

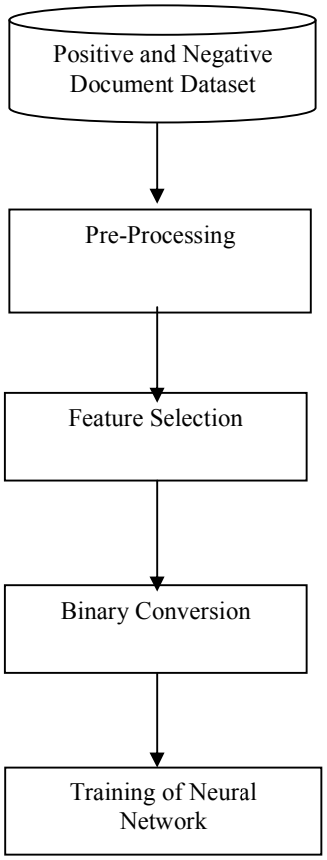

Fig.1 Proposed work training module.

Text preprocessing is consisting of words which are responsible for lowering the performance of learning models. Data preprocessing reduces the size of the input text documents significantly. It involves activities like sentence boundary determination, natural language specific stopword elimination. Stopwords are functional words which occur frequently in the language of the text (for example a, the, an, of etc. in English language), so that they are not useful for classification.

3.2 Feature Selection

Term

The vector which contain the pre-processed data is use for collecting feature of that document. This is done by

comparing the vector with vector KEY (collection of keywords) of the ontology of different area. So the refined vector will act as the feature vector for that document.
So the list of words which are crossing the threshold are consider as the keywords or feature of that document.

[feature] = mini_threshold ([processed_text])

In this way term feature vector is created from the document.

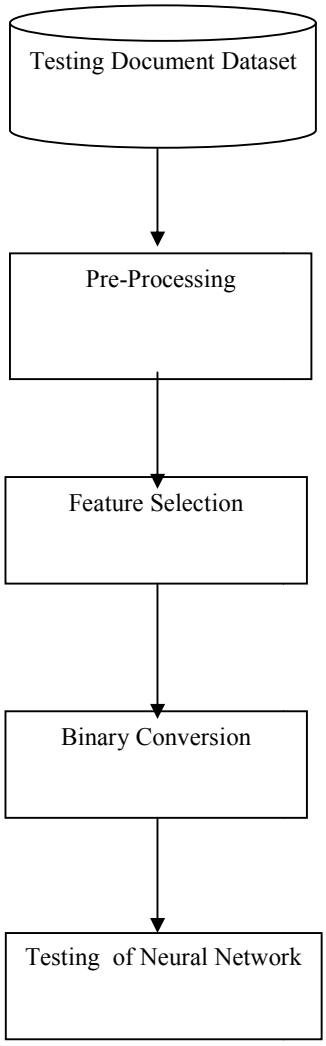

Fig. 2 Proposed work testing module.

3.3 Pattern

Here any consecutive term set is considered as the pattern in the document. As it is known that collection of patterns is performed in the separate set of features.

Positive and Negative Feature set

On the basis of work done in [8]. Classification of the terms is done in two category first is positive set of document and other is negative set of documents. In [8] two algorithms were proposed for classifying the terms into where some terms remain unclassified, so those terms are left in the work. In this way these vectors of positive and negative sets are considered as input in neural network for classification.

3.4 Binary Conversion

In this step keywords obtained from the features of the document are need to be insert into neural network for classification but as text words cannot be insert in the neural network. So a representative of those words are required. As each keyword is a set of ASCI value for example keyword "ABCD" ASCI set is [65 66 67 68]. Now each ASCI number is replace by its binary number as 65={ 1000001}, 66={ 1000010}, 67={ 1000011}, 68={ 1000100}. So in this work ABCD binary is {10000011000010100001111000100}.

As each word contain different number of characters so a set of 100 bit is taken as input in the neural network. Where default value is zero in the vector.

3.5 Training of Error Back Propagation Neural Network (EBPNN):

- Let us assume a four layer neural network.
- Now consider i as the input layer of the network. While j is consider as the hidden layer of the network. Finally k is consider as the output layer of the network.
- If w_{ij} represents a weight of the between nodes of different consecutive layers.
- So the output of the neural network is depend on the below equation:

$$Yj = \frac{1}{1+e^{-X_j}}$$

where, $X_j = \sum x_i \, . \, w_{ij} - \theta_j$, $1 \leq i \leq n$; n is the number of inputs to node j, and θ_j is threshold for node j

- The error of output neuron *k* after the activation of the network on the *n-th* training example (x(n), d(n)) is:

$$e_k(n) = d_k(n) - y_k(n)$$

- The network error is the sum of the squared errors of the output neurons:

$$E(n) = \sum e_k^2(n)$$

- The total mean squared error is the average of the network errors of the training examples.

- The Backprop weight update rule is based on the gradient descent method:

$$E_{AV} = \frac{1}{N} \sum_{n=1}^{N} E(n)$$

- It takes a step in the direction yielding the maximum decrease of the network error E.
- This direction is the opposite of the gradient of E.

- Iteration of the Backprop algorithm is usually terminated when the sum of squares of errors of the output values for all training data in an epoch is less than some threshold such as 0.01

$$w_{ij} = w_{ij} + \Delta w_{ij} \qquad \Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}}$$

3.6 Testing of EBPNN

In this step input query is preprocess as done in the training module, similarly feature vector is create by assigning identification numbers to those keywords. Finally feature vector is input in the EBPNN which give output. Now analysis of that output is done that whether specified class is desired one or not.

IV. Experiment and Results

In order to implement above algorithm for intrusion detection system MATLAB is use, where dataset is use of different size. Neural Network Toolbox includes command-line functions and apps for creating, training, and simulating neural networks. This make it easy to develop neural networks for tasks such as data-fitting, pattern recognition, and clustering. After creating networks in these tools, it can automatically generate MATLAB code to capture work and automate tasks.

4.1 Evaluation Parameter

As various techniques evolve different steps of working for classifying document into appropriate category. So it is highly required that proposed techniques or existing work need to be compare on same dataset. But document cluster which are obtained as output is need to be evaluate on the function or formula. So following are some of the evaluation formula which help to judge the classification techniques ranking.

Precision=True positive/ (True positive + False positives)

Recall=True positives/ (True positive + False negative)

F-Measure= 2*Precision*Recall/ (Precision + Recall)

Accuracy = (True Positive + True Negative) / (True Positive + True Negative+ False Positive + False Negative)

4.2 Dataset Description

In order to test proposed work performance testing is performed on real as well as artificial dataset. Here testing was done on four different sets of documents, named as D1, D2, D3, D4 where size of documents in these sets are 6, 8, 10, 11.

4.3 Results

Table 1. Precision and Recall testing results from trained Neural Network keyword class.

Dataset	Precision of Keyword Classification values on Different Testing dataset	
	Previous Work[8]	Proposed Work
90	0.2857	0.6667
80	0.3	0.6538
70	0.4615	0.6538
60	0.48	0.6400

Table 1 shows that proposed work has achieved a high precision value as the testing files are increasing. It has shown in table that trained neural network generated value is acceptable for the true positive case.

Table 2. F-Measure testing results from trained Neural Network keyword class.

Dataset	F-Measure of Keyword Classification values on Different Testing dataset	
	Previous Work[8]	Proposed Work
90	0.3636	0.5714
80	0.3750	0.5667
70	0.48	0.5667
60	0.4898	0.5614

Table 2 shows that proposed work has achieved a high recall value as the testing files are increasing. It has shown in table that trained neural network generated value is acceptable for the keyword classification. Accuracy can

© 2018 IJSRET

further be increased by passing high quality training dataset.

Table 3. F-Measure testing results from trained Neural Network keyword class.

Dataset	Execution Time (second) of Keyword Classification values on Different Testing dataset	
	Previous Work[8]	Proposed Work
90	0.045	0.0392
80	0.0614	0.0175
70	0.0365	0.0175
60	0.0231	0.0658

Table 3 shows that proposed work has achieved a low execution time value as the testing files are increasing. It has shown in table that trained neural network generated value in almost constant time, as compare to the previous work.

Table 4. Document classification Accuracy of testing results from trained Neural Network keyword class.

Dataset	Document Classification Accuracy values on Different Testing dataset	
	Previous Work[8]	Proposed Work
90	68.75	81.8182
80	62.5	75
70	50	75
60	43.75	71.4286

Table 4 shows that proposed work has achieved a high accuracy value as the testing files are increasing. It has shown in table that trained neural network generated value is acceptable for the keyword classification. Accuracy can further be increased by passing high quality training dataset.

V. CONCLUSION

With the extreme increment of the digital information on the servers or libraries it is critical for scientist to deal with it. Considering this work which has concentrated on one of the issue of document grouping which can be utilized by the distinctive association, for example, news, face off regarding, online articles, and so on. Here numerous researcher officially done part of work yet that is concentrate just on the content characterization where in this work document are arrange. In few work document classification are done on the basis of the background information, but this work overcome this dependency as well here it classify all the document without having prior knowledge. Results shows that using an correct iteration with fix number of neurons classification of keywords and documents is perform in proposed algorithm. As there is always work remaining in every because research is a never ending process, here one can implement similar thing for different other language.

REFERENCES

[1]. Dr. B. Poorna, Sudha Ramkumar. "Text Document Clustering Using Dimension Reduction Technique". International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 7 (2016) pp 4770-4774.

[2]. K. Fragos, P.Belsis, and C. Skourlas, "Combining Probabilistic Classifiers for Text Classification",Procedia - Social and Behavioral Sciences, Volume 147 Pages 307–312, 3rd International Conference on Integrated Information(IC-ININFO), doi: 10.1016 /j.sbspro .2014.07. 098 , 2014.

[3]. S. Keretna, C. P. Lim and D. Creighton, "Classification Ensemble to Improve Medical Named Entity Recognition", 2014 IEEE International Conference on Systems, Man, and Cybernetics, San Diego, CA, USA, 2014.

[4]. S.Ramasundaram, "NGramsSA Algorithm for Text Categorization", International Journal of Information Technology & Computer Science (IJITCS), Volume 13, Issue No : 1, ppp.36-44, 2014.

[5]. Jian Ma, Wei Xu, Yong-Hong Sun, Efraim Turban, Shouyang Wang, And Ou Liu. "An Ontology-Based Text-Mining Method To Cluster Proposals For Research Project Selection". Ieee Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans, Vol. 42, No. 3, May 2012

[6]. Disputant Relation-Based Classification For Contrasting Opposing Views Of Contentious News Issues Souneil Park, Jungil Kim, Kyung Soon Lee, And Junehwa Song. Ieee Transactions

On Knowledge And Data Engineering, Vol. 25, No. 12, December 2013.

[7]. Jian Ma, Wei Xu, Yong-Hong Sun, Efraim Turban, Shouyang Wang, And Ou Liu. "An Ontology-Based Text-Mining Method To Cluster Proposals For Research Project Selection". Ieee Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans, Vol. 42, No. 3, May 2012

[8]. Yuefeng Li, Abdulmohsen Algarni, Mubarak Albathan, Yan Shen, and Moch Arif Bijaksana. "Relevance Feature Discovery for Text Mining". IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 6, JUNE 2015..

[9]. S. Park, K.S. Lee, And J. Song, "Contrasting Opposing Views Of Contentious Issues," Proc. 49th Ann. Meeting Assoc. Computational Linguistics (Acl '11), Pp. 340-349, 2011.

© 2018 IJSRET