International Journal of Current Trends in Engineering & Technology
ISSN: 2395-3152
Volume: 03, Issue: 06 (NOVEMBER - DECEMBER, 2017)

Document Classification of Text Mining by Utilizing term and Pattern Features in EBPNN

Suraj Prasad, Prof. Manaswini Panigrahi
Department of computer Science & Engineering
surajprasadmit@gmail.com
+91-9870272160

Abstract— *As the internet users are increasing digital data on servers are increasing, this attracts researcher from text mining field to optimize various services. As various issues are arise on the server such as data handling, security, maintenance, etc. In this paper text classification is proposed that classify the text document in efficient manner. Here Error back propagation algorithm is utilizes for the classification which is a soft computing approach. Proposed classification approach classifies the data on the basis of initial learning where training of neural network is performing by binary input of the text vector. While in testing phase text documents are classify as per neural network training. Experiment is performing on real as well as artificial dataset. Result shows that proposed work is better as compared to previous work on different evaluation 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 contains 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. Content mining and information mining are comparative, aside from information mining deals with organized information while content mining takes a shot at semi-organized and unstructured information [9]. Information digging is in charge of extraction of certain, obscure and potential information and content digging is in charge of unequivocally expressed information in the given content. Then again potential data extraction is normal to both [20]. In this paper content archive mining calculation is proposed which are a blend of neural system based grouping calculations and other information mining strategies. Here terms are ordered first at that point reports have been grouped into the most fitting bunches, under which they have a place generally properly. The utilization of such content record mining strategies can be connected in dataset administration, to keep up information quality. This work can be connected so as to bunch and order the expansive number of reports that is chaotic. This is predominantly required for simple access to the precise record in least time. Therefore, enhancing the mining procedures that can be utilized as a part of the consistently developing size of reports gathered.

II. RELATED WORK

In this area few research work of this field is clarified which determine distinctive methodologies of the scientists. Here it is discovered that arrangement of content record required great arrangement of highlights for getting viable yield.

Dr. B. Poorna, Sudha Ramkumar in [1] has done content record bunching where gathering for an arrangement of reports was done in light of the data it contains and to give recovery comes about when a client peruses the web. In this work comes about demonstrates that proposed work has recover the content report productively by earlier arrangement of the content records in the archive.

Here work has concentrate on diminishing the measurement of the dataset. So measurement lessening is done in by two methodologies initially is decreasing of commotion or content which don't give any data while second is expelling of undesirable highlights from the record dataset.

K. Fragos et al. in [2] likewise finishes up for consolidating distinctive methodologies for content order. The strategies that creators have consolidated have a place with same worldview – probabilistic. Gullible Bayes and Maximum entropy classifiers are tried on the applications where the individual execution is great. The combining administrators are utilized over the individual outcomes. Greatest and Harmonic mean administrators have been utilized and the execution of mix is superior to the individual classifiers.

S. Keretna et al. [3] have taken a shot at perceiving named substances from a restorative dataset containing casual and unstructured content. For this, they consolidate the individual consequences of Conditional Random Field (CRF) classifiers and Maximum Entropy (ME) classifiers on the therapeutic content; every classifier prepared utilizing an alternate arrangement of highlights. CRF focuses on the logical highlights and ME focuses on the etymological highlights of each word. The joined outcomes were superior to the individual consequences of both the classifiers in view of Recall rate execution measure.

S. Ramasundaram et al. [4] planned to enhance the N-grams grouping calculation by applying Simulated Annealing (SA) seek strategy to the classifier. The cross breed classifier NGrams SA achieved an impromptu creation to the first NGrams classifier while acquiring every one of the upsides of Ngrams approach. Highlight lessening utilizing technique is utilized yet its multivariate incentive among the n-grams influences the execution of the classifier.

III. PROPOSED METHODOLOGY

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 organization for the information, for example, speaker's distinguishing proof image or exceptional character, here all procedure is perform by using the distinctive mix of terms and example highlights.

Preprocessing:- Preprocessing is a procedure utilized for change of record into include vector. Much the same as content arrangements the preprocessing likewise has debate about its division [1, 7]. Content preprocessing comprises of words which are in charge of bringing down the execution of learning models. Information preprocessing decreases the measure of the information content reports altogether. It includes exercises like sentence limit assurance, characteristic dialect particular stop word disposal. Stop words are useful words which happen every now and again in the dialect of the content (for instance a, the, an, of and so forth in English dialect), with the goal that they are not valuable for order.

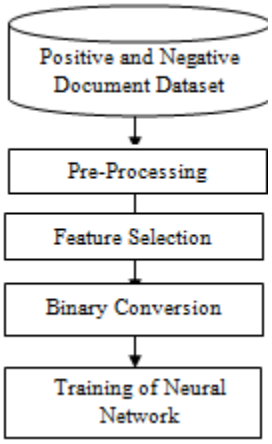

Fig.1 Proposed work training module.

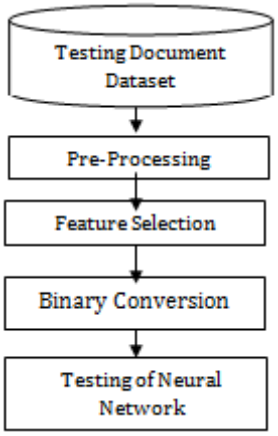

Fig. 2 Proposed work testing module.

Feature Selection:- The vector which contains the pre-prepared information is use for gathering highlight of that archive. This is finished by contrasting the vector and vector KEY (accumulation of watchwords) of the metaphysics of various territories. So the refined vector will go about as the element vector for that report. So the lists of words which are crossing the threshold are considered as the keywords or feature of that document. [Feature] = mini_threshold ([processed_text]), In this way term feature vector is created from the document.

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.

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 is 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 {10000011000010100000111000100}. As each word contains 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.

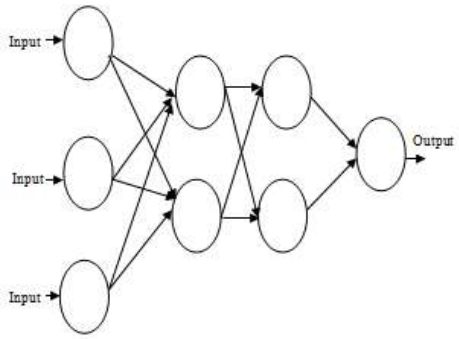

Fig. 3 Network activation Forward Step, Error propagation Backward Step

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 considered 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, $X_j = \sum x_i . Wij - \theta_j$, $1 \leq i \leq n$; 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: $_k(n) = d_k(n) - y_k(n)$ 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: -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

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

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

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

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 named as precision, recall, accuracy.

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.

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

Keyword Classification values on Different Testing dataset	
Precision	**Recall**
0.56	0.50
0.5556	0.50
0.625	0.50
0.6667	0.50

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 and Accuracy testing results from trained Neural Network keyword class.

Keyword Classification values on Different Testing dataset	
F-Measure	Accuracy
0.528	60
0.5263	59.259
0.5556	65.625
0.5714	69.0476

Table 2 shows that proposed work has achieved a high F-Measure and 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.

Table 3. . Accuracy of Document classification from trained Neural Network keyword class.

Document classification class Accuracy
Accuracy
50
50
50
72.7273

Table 3 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 drastic increase of the digital text data on the servers, libraries it is important for researcher to work on it. Considering this fact work has focused on one of the issue of document classification which can be used by the different organization such as news, debate, online articles, etc. Here many researchers already done lot of work but that is focus only on the content classification where in this work document are classify. 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 the entire document without having prior knowledge. Results shows that using a correct iteration with fix number of neurons classification of keywords and documents are 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.

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