International Journal of Technology Research and Management
ISSN (Online): 2348-9006
Vol 5 Issue 9 September 2018

A Neural Network Based Document Classification and Fetching by Using Term Features

Rajeev Kumar[1], Durgesh Wadbude[2]
Research Scholar, M.Tech (CSE), MIT Bhopal, India[1]
MIT Bhopal, India[2]
rajeevniranjan999@gmail.com[1], durgeshsvits@gmail.com[2]

Abstract: *As the text content are increasing day by day with the growing digital world. Researchers are working in this field from last few decades. In this paper a neural system is proposed to distinguish the content in proficient way. Proposed document distinguishing approach discover the terms in the document. Here frequent terms are filter out where each terms is arranged in matrix. Now clustering of the terms of document is done in efficient manner by using hierarchal clustering. Neural Network and clustering algorithm help in identifying the document without any guidance. Output of the work provide an dictionary which will help in classify the documents or comment done by social users of any network. Experiment is done on real dataset. Proposed work is compare with previous approach and results shows that proposed work is better as compare to previous work on different evaluation parameters.*

Keywords: *Genetic Algorithum, Feature Extraction, Text Categorization, Clustering.*

1. INTRODUCTION

Given the immense measure of unstructured data accessible online today, there is much to be picked up from the advancement of robotized frameworks that can viably sort out and group this information, with the goal that it well be utilized by human clients seriously. While it well be helpful to arrange this sort of data as per its topic, characterizing it as per the author feelings, or report, can likewise give specialists, business pioneers, and approach producers with significant data running from rates of consumer loyalty to popular sentiment patterns.

Document analysis has drawn great interest in recent years because of the surge of subjective content (blog posts, movie and restaurant reviews, etc.) being created and shared by Internet users, and the scope of new applications enabled by understanding the documents embedded in that content. For example, extracting the document of a review can help provide succinct summaries to readers, and can be very useful in automatically generating recommendations for users. Document classification can also help determine the perspective of different sources of information, and yet another possible application would be the processing of answers to opinion questions. Specifically within the field of reviews, the numerical ratings that come with many of them enable us to categorize them into finer-grained scales than just positive or negative categories. This more extravagant data makes it conceivable to rank things or quantitatively think about sentiments of a few analysts, in this way permitting more nuanced examinations to be completed.

The testing perspective in content / document investigation is a feeling word which is considered as a positive in one circumstance might be considered as negative in another circumstance. The customary content preparing looks at that as a little change in two bits of substance has no adjustment in the noteworthiness or importance [1]. However, in record investigation a little change in two bits of substance has change in the noteworthiness or importance, consider Example "story is great" is unique in relation to "the story isn't great". The framework procedure it by examining one by one sentence at any given moment [3]. Be that as it may, online journals and twitter contains more casual sentences which client can comprehend and yet framework can't comprehend it. Think about case, "that film story was comparable to its past motion picture" is subject to past motion picture whose points of interest isn't accessible.

Another testing part of this issue appears to recognize it from conventional point based arrangement is that while

subjects are frequently distinguished by catchphrases alone, record can be communicated in a more unobtrusive way [2]. For instance, the sentence "How might anybody watch this Drama?" contains no single word that is clearly negative. Along these lines point based order can without much of a stretch justifiable at that point record. Along these lines, aside from showing our outcomes acquired by means of machine learning strategies, we additionally comprehend the issue to pick up a superior comprehension of how troublesome it is. Consider another case visual impact of motion picture were great however storyline was frightful this pass on both positive and negative significance individually.

2. RELATED WORK

In [7] displayed a Bayesian arrangement system for content classification utilizing class-particular qualities. Not at all like consistent methodologies of content order proposed technique picked a specific component subset in each class. Applying such class-subordinate qualities for order, a Baggenstoss' PDF Projection Theorem was taken after to reproduce PDFs from class-particular PDFs and build a Bayes characterization run the show. The significance of recommended approach is that component choice criteria, similar to: MD (Maximum Discrimination), IG (Information Gain) are incorporated effectively. Assessed the execution on a few real benchmark informational collection and contrasted and highlight choice approaches.The tests , they tried approach for surface arrangement on paired ongoing benchmarks : 20-Reuters and 20-Newgroups.

In [8] proposed a BI-LSTM (Bidirectional long here and now memory) system to engrave the short content arrangement with 2 settings. The short-content order is required in uses of content mining, particularly human services applications in short messages mean etymological equivocalness bound semantic articulation because of which customary methodologies neglects to catch genuine semantics of restricted words. In social insurance areas, the content incorporates rare words, in which because of absence of preparing information installing learning isn't simple. DNN (Deep neural system) is potential to help the execution according to their quality of portrayal limit. At first, a typical consideration system was embraced to manage arrange preparing with space information in word reference. Also, coordinate situations when information word reference is inaccessible. They introduced a multi-errand model to learn space information word reference and performing content arrangement assignment in parallel. They connected recommended method to existing medicinal services framework and only accessible ATIS dataset to show signs of improvement results.

In [9] reviewed the procedure of content grouping and existing calculations. Substantial measure of information is put away as e-archives. Content mining is a method of separating information from these records. Ordering content archives in particular number of pre-characterized classes is Text grouping. Its application comprises of email directing, spam sifting, dialect recognizable proof, assessment investigation, etc.

In [10] presented a fluffy rationale based method to understand content grouping. Information embedded in proposed demonstrate are separated from twitter's message. Internet based life offers a lot of information to examine human conduct. It was utilized to extricate data and ordering content. It's valuable to dissect the connection between human impacted occasions and online networking. A few fluffy tenets are planned and de-fuzzification strategies were consolidated to get wanted outcomes. Proposed procedure was contrasted with well known inquiry strategy according to rate and amount accuracy. Results demonstrates that proposed procedure is appropriate for grouping of twitter messages. The trial utilizes the twitter survey utilizing web based life.

In [10] proposed an enhanced KNN content characterization calculation relying upon Simhash and normal Hamming hole of neighboring writings as a thing that tackles the issues created by information awkwardness and the extensive computational overhead in the customary KNN content grouping calculations. Tested outcomes clarified that proposed calculation performs with higher exactness and review and better F1 esteem. The investigation utilizes the mailing dataset like electronic and web mail information.

In [11] proposed a strategy which utilizes the association between lexical things and marks before completing Latent Dirichlet Allocation (LDA) subject show. They adjusted parameters of SVM (Support Vector Machine) to find enhanced qualities by K-wrinkle cross endorsement. It's a marvelous test that fathoming high-estimation and substance sparsity issues in short substance game plan. Additionally, using piece SVM as classifier, we adequately mastermind named short Chinese substance reports. Differentiating and other two customary systems k-Nearest Neighbor and Decision Tree of short substance course of action, the exploratory results exhibit that our procedure defeats them on arrange precision, exactness, survey and F-measure.

3. PROPOSED WORK

As the mining is utilize in different type of data analysis so for the same all need to increase the different technique in the required area. So contributing the text mining is done in this work by the proposed method for clustering the document or articles in the group without having any prior knowledge of the patterns. In the propose work no need of any format for the input data such as speakers identification symbol or special character, here all process is done by utilizing the different combination of cluster center field.

Pre-Processing: As the dataset is a gathering of information which is rough and need to recover essential data which is productive for the work. Content preprocessing is comprising of words which are in charge of bringing down the execution of learning models. Information preprocessing lessens the span of the information content / documents essentially. It includes exercises like sentence limit assurance, common words processing, stop word disposal and stemming. 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 helpful for grouping. Here work read entire undertaking and put all words in the vector. Presently again read the record which contain stop words at that point expel comparable words from the vector. Once the information is pre-process then it will be the accumulation of the words that might be in the dictionary.

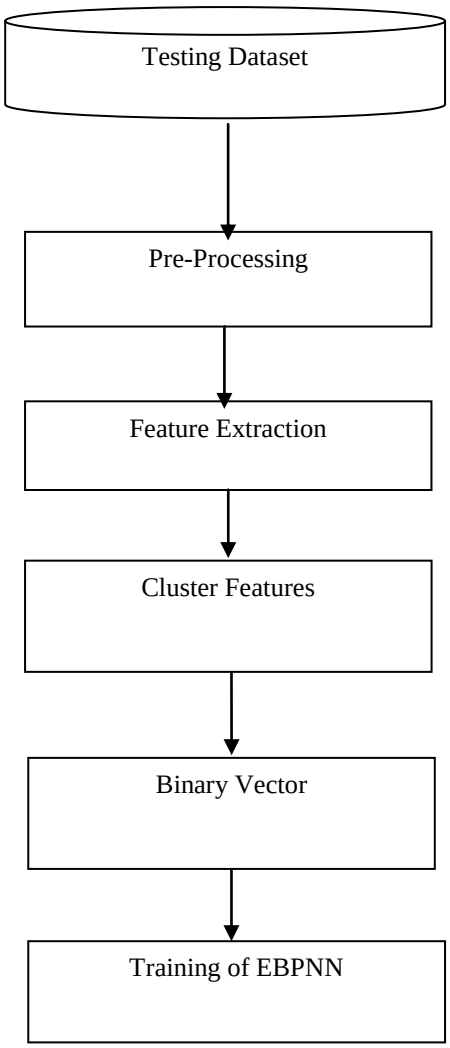


Fig. 1 Represent Block Diagram of Proposed work for Training.

For instance let one of the content / document class is taken, let the stop words gathering is S[] = {t1,t2,t3,… … .tm} and its content vector is Cd[] ={t1, ff1, ss1, t2, ss2, t3, t4, ff2… … ..tn}. At that point the vector acquire after the Pre-Processing is PP[] = {ff1, ss1, ss2, ff2,… … .ffx}.

Feature Keyword: Here keywords are selected from the H matrix which is a collection of the words after preprocessing from each document. In order to decide the word as a keyword, Support value has been decided so the words which

are crossing that limit is consider as keyword and put in the matrix of keyword K.

Cluster Features

In this step terms obtained from the documents are clustered. This clustering was so done that each cluster have different number of documents but as per term similarity from the cluster center grouping was done.

Let twenty documents are taken as input in the dataset and each document has n number of terms. Now first document act as cluster center make one cluster, after this next document terms are compared with first cluster terms if number of match terms cross threshold β than assign this document to the first cluster. While if matching term do not cross this threshold β than make new cluster for this second document. In similar fashion other set of documents are clustered.

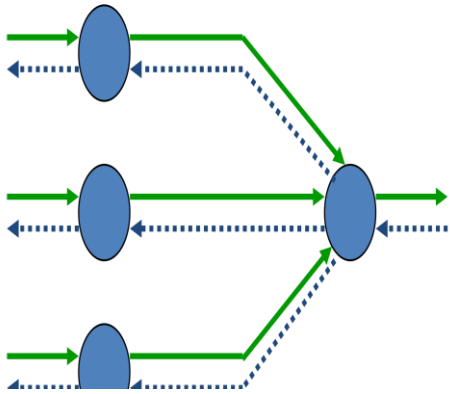
Binary Conversion

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 [70 71 72 73]. Now each ASCI number is replace by its binary number as 70={ 1000110}, 71={ 1000111}, 72={ 1001000}, 73={ 1001001}. So in this work ABCD binary is {1000110100011110010001001001}.

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.

Feature Vector: In this step keywords from the different query is put in a vector by its representative number. Here each vector segment is predefined. This can be understand as let V = [k1, k2, k3……..kn] then each keyword is identified by a unique number so vector feature vector is like [1, 2, 3, ……..11]and its segment is Xi. So this vector act as the input to the neural network while Xi act as the desired output.

Training of Error Back Propagation Neural Network (EBPNN):



Network activation Forward Step, *Error propagation Backward Step*

- Consider a network of three layers.
- Let us use i to represent nodes in input layer, j to represent nodes in hidden layer and k represent nodes in output layer.
- w_{ij} refers to weight of connection between a node in input layer and node in hidden layer.
- The following equation is used to derive the output value Yj of node j

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

where, $X_j = \sum x_i \cdot w_{ij} - \theta_j$, $1 \leq i \leq n$; 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.

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

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

$$w_{ij} = w_{ij} + \Delta w_{ij}$$

Proposed Algorithm
Input: D // Dataset, SN // Document Number

Output: SCD // Document cluster dataset

1. PD←Pre_Process(D) // Preprocessed Dataset
2. FP←Fetch_Keywords(PD) // FP: Fetch Pattern
3. Loop n=1: FP // Each Keyword
4. Loop m=1:Clust // m: Number of cluster
5. Match_Term[m]←Matching(FP[n], Clust[m])
6. If Match_Term[m]>β
7. Clust[m]←FP[n]
8. Break Loop
9. endif
10. EndLoop
11. EndLoop
12. Loop n=1: FP // Each Keyword
13. Bin_FP←Binary_Conversion(FP[n]) //
14. EndLoop
15. Trained_Network←Train(Bin_FP)

Testing of Trained Model

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.

4. EXPERIMENT AND RESULTS

Keeping in mind the end goal to execute above calculation for document classification framework MATLAB is utilize, where dataset was utilization of various size. Neural Network Toolbox incorporates number of inbuilt functions and applications for making, preparing, and recreating neural systems. This makes it simple to create neural systems for work, for example, information fitting, design acknowledgment, and bunching. Here dataset of document class is pass in random fashion for increasing the testing complexity of the work.

a. 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: Precision value is the ratio of predicted positive user to the totalpredicted user.

$$Precision = \left(\frac{True_{positive}}{(False_{positive} + True_{positive})} \right)$$

Recall: The recall is the fraction of relevant users that have been predicted over the total amount of input users. It is also known as Sensitivity or Completeness.

$$Recall = \left(\frac{True_{positive}}{False_{negative} + True_{positi\ ve}} \right)$$

F-Measure: Harmonic mean of precision value and recall value is F-measure.

$$F - Measure = \left(\frac{2xPrecisionxRecall}{(Recall + Precision)}\right)$$

Accuracy: This act as the percentage of correct prediction from the total set of prediction.

$$Accuracy = \left(\frac{Correct\_class}{(Correct\_class + InCorrect\_class)}\right)$$

Results

Results of the proposed work is compare with the existing method in [12].

Table 1: Precision value comparison of proposed and previous work.

Emotion	Precision Value Comparison	
	Proposed Work	Previous Work
Query1	0.642857	0.214286
Query2	0.5	0.214286
Query3	0.571429	0.285714

Above table 1shows that precision value of proposed work was high as compared to previous work [12]. It has been observed that proposed work trained neural network method is efficient as compare to the previous. Here iteration in both work increase the precision value but selection of perfect set of features for clustering make high precision value of proposed work.

Table 2: Recall value comparison of proposed and previous work.

Emotion	Recall Value Comparison	
	Proposed Work	Previous Work
Query1	0.692308	0.230769
Query2	0.538462	0.230769
Query3	0.615385	0.307692

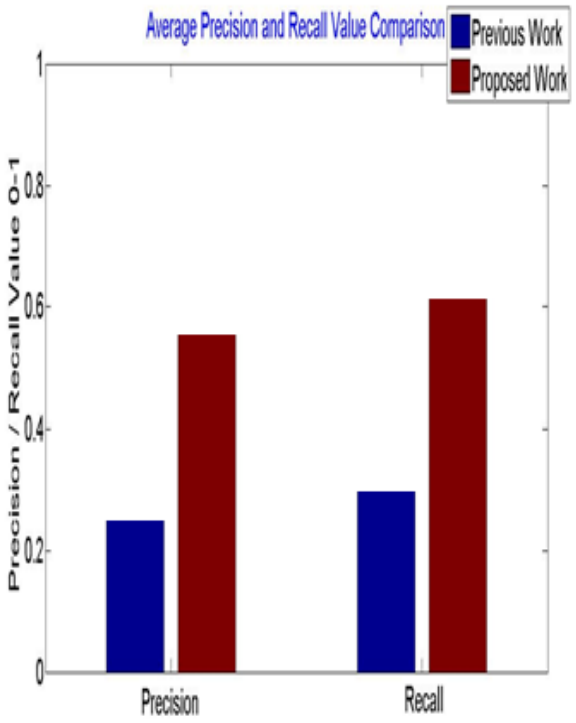


Fig. 2: Average Precision and Recall Value comparison.

Above table 2 shows that recall value of proposed work was high as compared to previous work[12]. It has been observed that proposed work trained neural network method is efficient as compare to the previous. Here iteration in both work increase the precision value.

Table 3: Accuracy value comparison of proposed and previous work.

Emotion	Accuracy Value Comparison	
	Proposed Work	Previous Work
Query1	0.9	0.6
Query2	0.7	0.5
Query3	0.8	0.8

Above table 3 shows that f-measure value of proposed work was high as compared to previous work[12]. It has been observed that proposed work trained neural network method is efficient as compare to the previous. Here iteration in both work increase the precision value but selection different set of features for clustering make high accuarcy value of proposed work.

Table 4: Accuracy value comparison of proposed and previous work.

Emotion	Execution Time in second Comparison	
	Proposed Work	Previous Work
Query1	0.0166775	0.135585
Query2	0.0349018	0.133032
Query3	0.0276762	0.134613

5. 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 focus on one of the issue of the clustering based document fetching which is build by the different organization such as news, debate, online articles, etc. Here many researchers has already done lot of work but that is focus only on the document custering where in this work terms are classify. In few work term clustering 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. By using trained neural network EBPNN, clustered data is easy t transfer and configure. Results shows that using an correct iteration with fix proposed algorithm works better than previous work. 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] Andreea Salinca. “Convolutional Neural Networks for Document Classification on Business Reviews”. arXiv:1710.05978v1 [cs.CL] 16 Oct 2017.
[2] Vishwanath Bijalwan, Vinay Kumar, Pinki Kumari, Jordan Pascual. “KNN Based Machine Learning Approach For Text And Document Mining”, 2014,Vol.7,No.1,Pp.61- 70.
[3] Tanmay Basu, C. A. Murthy, “Effective Text Classification By A Supervised Feature Selection Approach”,2008.
[4] Gautami Tripathi and Naganna S, “Feature Selection and classification approcha for Document Analysis”, Machine Learning and Applications: An International Journal (MLAIJ) Vol.2, No.2, June 2015
[5] Hemalatha1, Dr. G. P Saradhi Varma, Dr. A.Govardhan,“Document Analysis Tool using Machine Learning Algorithms “,International Journal of Emerging Trends & Technology in Computer Science Volume 2, Issue 2, March – April 2013
[6] Anurag Mulkalwar, Kavita Kelkar Document “Analysis on Movie Reviews Based on Combined Approach”, International Journal of Science and Research, Volume 3 Issue 7, July 2014
[7] B.Tang, H. He, et al., "A Bayesian classification approach using class-specific features for text categorization." IEEE Transactions on Knowledge and Data Engineering 28, pp: 1602-1606,no. 6, 2016.
[8] S. Cao, B. Qian, et al.," Knowledge Guided Short-Text Classification for Healthcare Applications", 2017 IEEE International Conference on Data Mining (ICDM) vol. 2, no. 6,pp: 234-289. 2017.
[9] V. K. Vijayan, K. R. Bindu, et al., "A comprehensive study of text classification algorithms." IEEE Advances in Computing, Communications and Informatics (ICACCI),, vol 12, no. 1 pp: 42-53. 2017.
[10] J. Liu, T. Jin, et al., "An improved KNN text classification algorithm based on Simhash." In Cognitive Informatics & Cognitive Computing (ICCI* CC), 2017 IEEE 16th International Conference on, pp. 92-95. IEEE, 2017.
[11] X. Wang, J. Wang, et al., "Labelled LDA-Kernel SVM: A Short Chinese Text Supervised Classification Based on Sina Weibo." In 2017 4th International Conference on Information Science and Control Engineering (ICISCE), pp. 428-432. IEEE, 2017.
[12] Chi Chen, Xiaojie Zhu, Student Member, IEEE, Peisong Shen, Jiankun Hu, Song Guo, Zahir Tari, AndAlbert Y. Zomaya. “An Efficient Privacy-Preserving Ranked Keyword Search Method”. IEEE Transactions On Parallel And Distributed Systems, Vol. 27, NO. 4, APRIL 2016
[13] Dandan Jiang, Xiangfeng Luo, Junyu Xuan, And Zheng Xu. “Document Computing for the News Event Based on the Social Media Big Data”. Digital Object Identifier 0.1109/ACCESS.2016.2607218, IEEE Access March 15, 2017.