Text Classification Using Fuzzy Neural Network

U. Sree Krishna, Hima Shree, K. Jayadeep, P.Lakshmi Prasanna

Abstract— In today's world, Large documents are being produced every day which require to be organized and also have the ability to extract the data of out of them. This organization into various topics is known as text classification. To perform text classification, many efficient algorithms are available, but this paper will focus on Text classification using Fuzzy Neural Networks. The First step in the algorithm that we chose is preprocessing. In this step, all the words are divided into tokens and stop words are removed along with the stemming also being done. The next step in this process in feature extraction, this process selects a subset of key words which best represent the text documents to be able to classify the document properly. The process however cannot yield 100% accuracy but has been refined in the modern-day world up to 94%.

Keywords: Text classification, Neural networks, Fuzzy, documents, Decision Tree.

1. INTRODUCTION

Daily, in media and various applications, we have a lot of information that is being created which is mostly similar. All these types of data are createdwhen they touch with others such as sharing idea, or proposing new on some sort of general topic. This text is considered to be unstructured data that possesses traits such as sparsity, ambiguity, and dimensionality. Text classification for documents is based upon some existing defined category. Text classification algorithms serve the purpose of classifying text documents which have predefined classes, many of these supervised classifications algorithms include Decision Tree Algorithm, Naive Bayesian, and few other algorithms [3].

Decision tree algorithm is a supervised learning algorithm which provides the easiest way to represent data when compared to other algorithms, it is used to classify certain sets of data. A decision tree creates a training which predicts class or target value using a set of decision rules [9].

On the other hand, Naive Bayes classifier is based on the Bayesian theorem and it is used when dimensionality is high in range, it is used for calculating possible output based on the data. It adds new raw data at runtime and has good classifier [4].

We also have neural network algorithm, in which we provide a fixed input to a neural network layer that acts to something similar to that off a brain neuron, it takes the input of worded texts or documents, give each word in the document a weight to consider its relevance to the topic. The artificial neural network is an information processing pattern that works in ways similar to the nervous system. It consists of a high number of interconnected nodes that work together to obtain the result of a problem. In this paper, We'll be using convolutional neural networks in order to classify our document.

Revised Version Manuscript Received on March 08, 2019.

U. Sree Krishna, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India

Hima Shree, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India

K. Jayadeep, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India

P.Lakshmi Prasanna, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India

2. FUZZY LOGIC

Fuzzy logic is an approach we use for computing the logic or the "degree of truth" based upon on where it is true or false (1 or 0) values are assigned to text in fuzzy logic. Fuzzy logic is closer to in the way our brain work, we take the data and form a partial truth for which further scope on for higher truth then if a certain threshold is exceeded, for those certain results, we get a motor reaction.

3. CONVOLUTIONAL NEURAL NETWORK

CNNs are like neural networks, they are made up of neurons that have weights that can learn. Every single neuron receives several inputs [19], and then it takes a weighted sum for each neuron, and passes it to an activation function which finally gives the output [15]. In this, The CNN has a loss function and the preparation of neural networks we made here can be applied for CNN as well. [11].

4. EXISTING SYSTEM

The existing system consist of a plain old neural network which consist of the basics, which are the input layer, hidden layers, and the output layer. Like the basic neural networks, the job of each layer works accordingly but there is no sense of fuzziness in the terms of the classification of the data. The input layer will take a sentence as the input while the hidden layer will calculate the weights at a precise level and the output layer will display precisely to which class it belongs to [18]. What's wrong the existing system? Well, sometimes this tends to misclassify the specific documents which causes a lower accuracy when compared to other classification methods such as SVM, and Naïve Bayes. ANN is a dynamic system that can change it structure with respective to the external or internal information that is fed through the network[12]. The existing system does not allow for approximations which can sometimes become a drawback when the classifier is unsure on where to classify certain sets of data.

5. PROPOSED SYSTEM

In this report, the method that we propose will consist of four main phases, which are text processing, feature extraction, text classification using MLF algorithm, and finally the evaluation of the results. Initially, the text preprocessing is responsible for diving each individual word in the document into the terms(tokens) [17]. Next the feature extraction occurs to remove the unnecessary components of the texts, the specific vector space is then moved to another dimension for the newly created vector space to remove the less important dimensions [20]. Finally, The MLF algorithm is applied to determine the classes of each of the sentences [10]. The results are then evaluated on the basis of how accurately they were determined. So, what differs in our proposed methodology from the original? The main idea in our methodology is to add fuzziness to the already given neural networks and further enhance its classification capabilities along with providing a faster performance than previous methods. So, the main point of the proposed system is to add fuzziness and to classify data as per the given needs. This will all be performed in the matter of within a minute which allows this proposed system to stand out from the other classifying methodologies.

6. LITERATURE SURVEY

S. no	Author	Title	Methodology	result	comparison	datasets	Existing Method	Tear of Publication
1	Amir KaramiAryyaGangopadhyayBinZhouHadi Kharrazi	Fuzzy Approach Topic Discovery in Health and Medical Corpora	The methodology in thispaper consists of using a fuzzylatent semantic analysis in order to handle health & medical redundancy. This helps solve a numerous problem in the field		medical and Health			2013
2	Rene Witte1 and Sabine Bergler2	Fuzzy Clustering for Topic Analysis and Summarization of Document Collections	A fuzzy clustering algorithm is used to to analyze collections of documents that could have resulsted from any query on an document server	Ten documents were collected and based on the given question, the cluster graph was formed				
3	SubhasreeBasu∗, Yi Yu† , Roger Zimmermann∗	Fuzzy Clustering of Lecture Videos Based on Topic Modeling	In this, The use of LDA in fuzzy clusttering to group together multimedia files such a tutorial videos was attemepted. It was grouped together based on the similar keywords		Fuzzy C-Means 0.453, PLSA,k-Means		The video were collected from youtube channel forNational Programme on Technology Enhanced Learning(NPTEL)	

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

4	RubayyiAlghamdi ,KhalidAlfalqi	A Survey of Topic Modeling in Text Mining	In this, LSA and LDA were used to for correlated topic modeling	Observed	Survey on all 4 methods			2015
5	Yu Chen∗, Rhaad M. Rabbani∗, Aparna Gupta†, and Mohammed J. Zaki		Text analytics were calculated based upon topic modeling in banking	1 day 0.7 0.65 0.54 0.39	NMF, PCA, LDA and KATEfor the 8-K and 10-K data, respectivel y	8-K and 10-K filings, from the years 2005–2 016	These methods include Principal Componen t Analysis, Non-negati ve Matrix Factorizati on, Latent DirichletAl location and KATE	2017
6	ZhenxingNiu,Gan gHua,LeWang,Xi nbo Gao.	Knowledge-Based Topic Model for Unsupervise d Object Discovery and Localization	LDA along with a combination of Dirichlet trees were used to integrate links into topic modeling for the discovery of objects	This method improves upon the topic coherence and is able to outperforms some unsupervised methods for the discovery of objects		Object discove ry, object localiza tion, latent Dirichl et allocati on.	LDA with must-links	2018
7			Ken Gorro1, Jeffrey Rosario Ancheta2, Kris Capao1, Nathaniel Oco2, Rachel Edita Roxas2, Mary Jane Sabellano1, Brandie Nonnecke3, Shrestha Mohanty3, Camille Crittenden3, and Ken Goldberg3	It involves the analysis of reduction in disaster risk suggestion with the help of topic modeling and as well as word2vec	The word2vec has a relatively high score which tells us that the words are closely related and it lets the community know how well to prepared in any event			

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

8	Anamta Sajid, Sadaqat Jan and Ibrar A. Shah	Automatic Topic Modeling for Single Document Short Texts	It involves in automizing the process of extracting topics of any given title of a document	Nouns are considered to be more reliable and better to finding certain topics of a given text	They are compared to find the most suited approach for extraction of a certain topic	relevance,novelty	topic modeling.	2017
9	Jennifer Sleeman, Milton Halem, Tim Finin, Mark Cane	Discovering Scientific Influence using Cross-Domain Dynamic Topic Modeling	It involves the uses of cross domainanalytics to aid the correlations between certain chapters and their documents	It is done by the predicting the importance of the extracted topic and then assessed through a form of cross domain correlation		cross-domain correlation; data integration; domain influence;	assessment reports of the Intergovernmental Panel on Climate Change (IPCC)	2017
10	Xiaoping Sun	Textual Document Clustering using Topic Models	Topic modeling	The method is achieved by comparing the accuracy of the latest topic modeling algorithms	We compare them with any other cluster topic modeling algorithm or methods	Document,probabilistic	TFIDF model	

7. THEORETICAL ANALYSIS/ARCHITECTURE

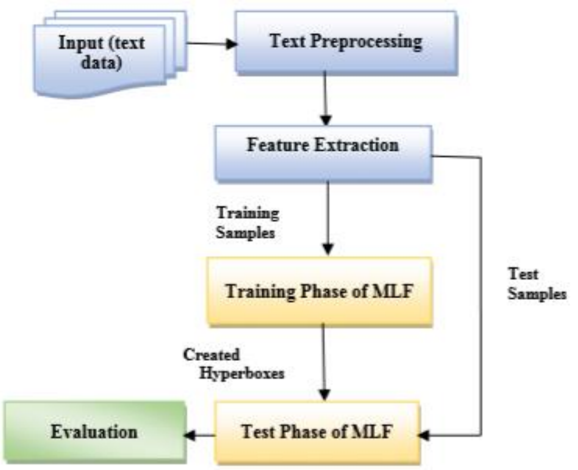


Figure 1

The basic architecture is still based upon the artificial neural networks in a way that all the layers are still present for the input, and the calculations along with the output [13]. The neural network architecture is already well established in the data science field. In this, we further enhance this architecture with the addition of a fuzzy logic features to the networks. This allows us to classify at higher accuracy than the stand-alone neural network.

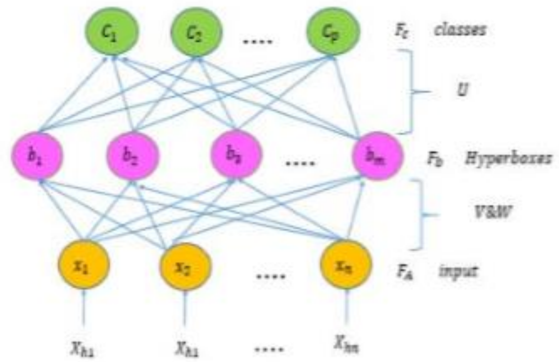


Figure 1.1

In the above Figure, the FMMN has 3 total layers which pertain to each task of a Neural network. The first layer is where the inputs are received, the second layer is the hidden layer which does the hyperbox clustering, and finally the third layer nodes represent a class for theclustering. All the nodes are interconnected through a series of links and each of them have a weight.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

8. ALGORITHM

Preprocessing

X_i=< X_{i1}, X_{i2},…………X_{in}>

= < P(C1/Wi), P(C2/Wi ……………., P(Cn/Wi >

for $i \le j \le p$.

In this, $d_{qi}$is meant to indicate the number of occurrences of w_i in document d_q

∂_{qj} can be defined between the values of either 1 or 0

Therefore, we have m word patterns in total.

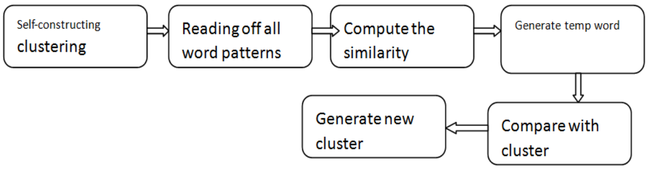


Figure 2

Algorithm Used

$$tfidf(W)=tf^*\log\frac{N}{df(W)}$$

Here, tf is meant to represent the number of times a word occurs in the documents, df(w) represents how many number of documents use that words and N is thetotal number of documents that we used

Membership Function:

$$b_j(X_h)=\frac{1}{2}\sum_{i=1}^{n}[max(0,1-max(0,\gamma\, min(1,X_{hi}-w_{ji})))+max(0,1-max(0,\gamma\, min(1,v_{ji}-X_{hi})))]$$

Here, X_h=(X_{h1},X_{h2},… X_{hn}) is the hth sample in the function and y is considered to be a coefficient that can adjust the total variable distance that lies from $X_h$and B_j . The two variables lie between he membership function values of 0 and 1. $V_j$and W_j are the minimum and maximum points of specific hyperboxB_j.

9. EXPERIMENT RESULTS

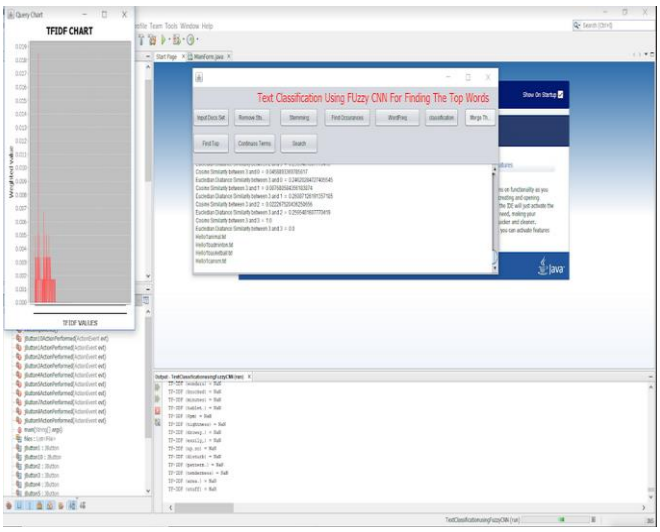


Figure 3

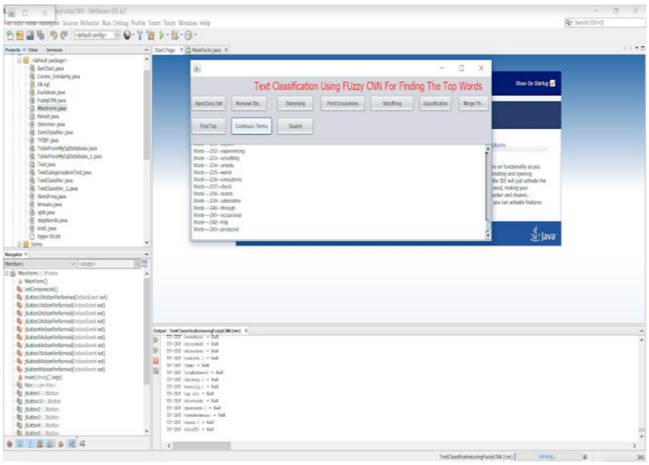


Figure 3.1

The experiment was performed in Java as we could integrate more text classification techniques which would help further with the methods. We used techniques such as stemming, TF-IDF as well to perform the experiment. The experiment was able to successfully classify the documents and is able properly check if the word is relevant to a section or not as that is the purpose of the text classifications. We also used SQL and K-means algorithm along with CNN due to some short text word relevancy [1].

10. DISCUSSION OF RESULTS

For this experiment, a newgroup corpus dataset was which consisted a set of newspapers. However, For the simplicity of project presentation, the code will be set for a few sentences in order to save run time and allow for a presentable timetable.

These are the following results from the MLF Classifier,

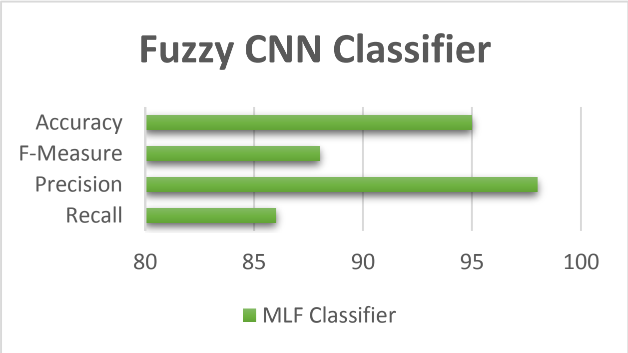
- In accuracy, the classifier had an accuracy of 95% of the newgroup dataset.
- In precision, the classifier had the precision of 90% on the newgroup dataset.
- In recall, the classifier had a recall of 86% on the newgroup dataset,
- In F-measure, the classifier has the F-measure of 88% on newsgroup dataset.
- This classifier also had the lowest running time among all the other classifiers used to compare.

Table 1

Classifier	Recall (%)	Precision(%)	F-Measure(%)	Accuracy(%)	Running Time
MLF Text Classifier	86	90	88	95	30.5

Graph 1

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication



11. CONCLUSION

In this research paper, the problem behind text classification was analyzed and solved using the means of Fuzzy Neural Networks. Initially, the document is tokenized and then the features are extracted [8]. The Multi-Level Fuzzy Network is then allowed to classify the documents as per the given data. We've been able to reach the conclusion that fuzzy neural networks are indeed accurate at classifying the data given up to the extent of 95%. Combined with the low runtime for this algorithm, this makes the MLF an good alternative to other methods such as Naïve Bayes and Support Vector Machine[2]. The topics were classified as per the training data topics and were mostly accurate. So, we can safely assume that the MLF is a good classifier and performs the tasks it was meant to do efficiently. This project can be furthered in the future with the help of tools such as TensorFlow which allow for a deeper and more integrated mining of text within a document to yield higher results. We would like to conclude by stating that MLF is one of the efficient ways for text classification given that our current era deals with a high number of large documents and that MLF is able to classify them with the utmost accuracy with the lowest runtime. Fuzzy neural networks is truly one of the keys to the future.

REFERENCES

1. Bao Y. and Ishii N., "Combining Multiple KNN Classifiers for Text Categorization by Reducts", LNCS 2534, 2002, pp. 340-347
2. Bi Y., Bell D., Wang H., Guo G., Greer K., "Combining Multiple Classifiers Using Dempster's Rule of Combination for Text Categorization", MDAI, 2004, 127-138.
3. Brank J.,GrobelnikM., Milic-FraylingN.,Mladenic D., "Interaction of Feature Selection Methods and Linear Classification Models",
4. W. Ding, S. Yu, Q. Wang, J. Yu and Q. Guo, "A Novel Naive BayesianText Classifier," in Information Processing (ISIP), May 2008.
5. C. Luoa, Y. Lib and S. M. Chungc, "Text document clustering based onneighbors," Data & Knowledge Engineering, vol. 68, no. 11, pp. 1271-1288, 2009.
6. Michael W. Berry, Survey of Text Mining: Clustering, Classification,and Retrieval, Amazon, 2003.
7. R Basili and A. Moschitti, "A robust model for intelligent textclassification," Proc. of ICTAI-01, 13th IEEE InternationalConference on Tools with Artificial Intelligence, IEEE ComputerSociety Press, Los, 2001, pp. 265-272.
8. Liao S. and Jiang M. "An Improved Method of FeatureSelection Based on Concept Attributes in TextClassification". Lecture Notes in Computer Science, Vol.3610, pp. 1140 – 1149, 2005.
9. Amasyali, M. F., &Ersoy, O. (2008). Cline: A new decision-tree family. IEEETransactions on Neural Networks, 19(2), 356–363.
10. T.Mikolov, I. Sutskever, K. Chen, G. S.Corrado, and J. Dean,"Distributed Representations of Words and Phrases and TheirCompositionality," Advances in Neural Information ProcessingSystems, 2013, pp. 3111–3119.
11. Zhan Wang, Yifan He and Minghu Jiang, "A Comparison among Three Neural Networks for Text Classification", Lab of Computational Linguistics, Dept. of Chinese Language, Tsinghua University, Beijing, 100084, China, 2006.
12. M.Ghiassi , M. Olschimke, B. Moon, P. Arnaudo, "Automated text classification using a dynamic artificial neural network model", Santa Clara University, 500 El Camino Real, Santa Clara, CA 95053-0388, USA, 2012.
13. Lea Vega and Andres Mendez-Vazquez, "Dynamic Neural Networks for Text Classification", Dept. of Electrical Engineering and Computer Science CINVESTAV Guadalajara Zapopan, México, 2016.
14. Jenq-Haur Wang, Hsin-Yang Wang, "Incremental Neural Network Construction for Text Classification", Department of Computer Science and Information Engineering, National Taipei University of Technology Taipei, Taiwan, 2014.
15. Jian-hai Du, "Automatic Text Classification Algorithm based on Gauss Improved Convolutional Neural Network", Journal of Computational Science, 2017.
16. Rodrigo Fernandes de Mello,Luciano Jose Senger, LaurenceTianruo Yang, "Automatic Text Classification Using AnArtificial Neural Network", 2002.
17. Cheng Hua Li and Soon Cheol Park, "Neural Network for Text Classification Based on Singular Value Decomposition", Division of Electronics and Information Engineering, Chonbuk National University Jeonju, Jeonbuk, 561-756, Korea, 2007.
18. DigantaSaha, "Web Text Classification Using a Neural Network", Department of Computer Science and Engineering Jadavpur University Jadavpur, Kolkata, India, 2011.
19. Lin Li, Linlong Xiao, Nanzhi Wang, Guocai Yang, School of Computer and Information Science Southwest University Chongqing, China, Jianwu Zhang, Dean's Office Tianjin Railway Technical and Vocational College Tianjin, China, "Text Classification Method Based on Convolution Neural Network", 2017.
20. ManabuNii, YuyaTsuchida, Yusuke Kato, Atsuko Uchinuno and Reiko Sakashita, "Nursing-care Text Classification using Word Vector Representation and Convolutional Neural Networks", Graduate School of Engineering, University of Hyogo, Himeji, Hyogo, Japan, 2017
21. Neustein, Amy, S. Sagar Imambi, et al. "Application of text mining to biomedical knowledge extraction: Analyzing clinical narratives and medical literature " De Gyter publication, 2014

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication