TEXT CLASSIFICATION BASED ON FUZZY RADIAL BASIS FUNCTION

Zuhair Hussein Ali [1]
[1] College of Education/ AL-Mustansiriyah University,
Baghdad, Iraq
zuhair72h@uomustansiriyah.edu.iq

Ameen A. Noor[2]
[2] College of Education/ AL-Mustansiriyah University,
Baghdad, Iraq
a.ameen63@uomustansiriyah.edu.iq

Abstract: ***Automated classification of text into predefined categories has always been considered as a vital method in the natural language processing field. In this paper new methods based on Radial Basis Function (RBF) and Fuzzy Radial Basis Function (FRBF) are used to solve the problem of text classification, where a set of features extracted for each sentence in the document collection these set of features introduced to FRBF and RBF to classify documents. Reuters 21578 dataset utilized for the purpose of text classification. The results showed the effectiveness of FRBF is better than RBF***

Keywords: Text classification, RBF, FRBF, Information gain, mutual information.

I. INTRODUCTION

According to the development of social networks, a massive amount of text data is rapidly created. The requirement for a well-characterized philosophy to investigate and classify these huge data has attracted many researchers to this kind of data which is known as unstructured data. Text classification (TC) is the solution to such problem of information overload [1]. TC is the process of assigning labels or categories to text according to its content. It's one of the important tasks of Natural Language Processing (NLP) with wide applications such as spam detection, sentiment analysis and topic labeling [2]. There are two ways for TC manually and automatically, manual classification can be done by a human that can provide a quality result, but it's expensive and time consuming, whereas the automatic classification applies NLP techniques and machine learning methods to classify text in faster and less costly [3].

TC models can be divided into supervised and unsupervised models. Supervised learning is a learning in which the trainer learns or teach the machine using well defined data. The supervised model consists of two phases training phase and testing phase. During the training phase a set of known labeled data feeds to the machine learning algorithm. The goal of this phase is to reach the desired output by train the algorithm. Through the testing phase, a set of unknown, labeled data feeds the algorithm to classify the data into classes depending on the training phase. Unsupervised learning is the training of the model using data that is neither classified nor labeled and allowing the model to act on that data without guidance [4]. Classification can also be divided into binary classification or multiclass classification. In binary classification the data is assigned to one of two classes, while in multiclass classification the data assigned to more than two classes based on the classification rules [5].

In this a method based on Radial Basis Function (RBF) and fuzzy Radial Basis Function (FRBF) introduced. Firstly, a set of features that include document frequency, TF_IDF, Mutual information and Information Gain are extracted from each sentence in the document collection, these features are used in RBF and FRBF for classification.

II. RELATED WORK

Over the most recent 20 years Information Retrieval (IR) and content-based document have picked up a noticeable status in the data organization field. It is because of the expanded accessibility of documents in digital form. TC is the ability to assign a label to a document that is very important in the field of IR. In this section some of TC methods will be investigated.

In 2017 Conneau et al. Build a text classification model based on the character's level. The model called Very Deep Convolutional Neural Networks (VDCNN) the model starts with a look –up table that creates a 2D tensor that contains a number of the embedding character. The model starts by applying one layer of 64 convolutions of size 3, followed by a stack of convolutional blocks. Each layer has the same number of feature map, also the model contains 3 pooling operations to reduce the memory size. The model proves the performance by using more than 29 convolutional layers [6].

In 2018 Rajshree Jodha et al. used K-Nearest Neighbors (KNN) for TC. In this model vector space model used to represent each document and each dimension of the vector correspond to distinct term, TF-IDF used to assign weight to each term in the document. Term weight used to train KNN which is supervised learning where each new document is compared to the train documents then the algorithm decided the class of the new document based on the training documents [7].

In 2019 Liang Yao et al. used Text Graph Convolution Networks (Text GCN) for TC. Text GCN basically builds on GCN which is semi supervised learning consists of a multilayer neural network that work on graph directly. Firstly, based on the document word relations and word co-occurrence a single graph text was built, then Text GCN can be learned. The learning of Text GCN is done by initializing it with one representation of word and document. The

process of learning will be expanded for more words and documents by building more graph where the number of nodes in the graph represents the number of documents in the corpus. Feature matrix that consists of Term Frequency-Inverse Document Frequency (TF-IDF) feed to Text GCN. Edge of graph builds based on feature matrix. Finally, learning for this feature done for classification purpose [8].

III. PROPOSE MODEL

The proposed method consists of many phases as shown in Figure (1). These phases include: preprocessing, feature extraction, then applying a Radial Basis function (RBF) and Fuzzy Radial Basis Function (FRBF).

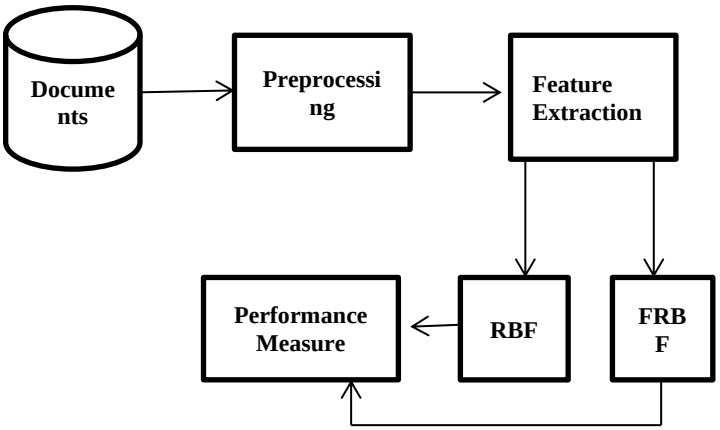

Fig.1: Block diagram of the Proposed Model.

A) Preprocessing

There are three steps for preparing the data, these steps are:

1-Tokenization: The main goal of tokenization is separate sentences into words.

2- Stop Words Removal: is the manner of removing words that appear many times in the text and don't offer the required information for recognizing an important sense of the document. There are many strategies utilized for indicating such stop words list. Now, various English stop word list is generally utilized to the TC procedure.

3- Stemming: is the method of generating origin of the word [9]. In this research Proter stemmer applied in this research [10].

B) Feature Extraction

Feature extraction is essential part in any TC method. The feature extraction based on computing term weight. Term weighting is the process of assigning a weight to every term in the document set based on the term importance in the document set. The features include document frequency, TF_IDF, Mutual information and Information Gain.

1- Document Frequency (DF)

Computes how many a term T_i appear in all documents. DF can be calculated as in Eq. (1)[11].

$$DF = \sum_{i=1}^{M} Ti \qquad (1)$$

Where M is the number of documents in the corpus.

T_i term frequency in document DF.

2- Mutual Information (MI)

The MI of two random variables tell us the quantity of dependence between these two variables. MI estimates how much information the existence/nonappearance of a term contributes to creating the right classification choice. MI can be calculated as in Eq. (2)[12].

$$I(X,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y) log_2 \frac{p(x,y)}{p(x)\,(y)} \qquad (2)$$

3-Term Frequency an Inverse Document Frequency (TF-IDF)

Identifies the term significance is very useful in TC system, that can be done by weighting term which can be calculated by multiplying TF by the IDF Which can be calculated as in Equation (3)[13].

$$TF - IDF = TF * \log\left(\frac{M}{df}\right) \qquad (3)$$

Where

TF = is the term frequency.

df = is the number of the document is which the term appears.

4- Information Gain (IG)

Measures how much a feature gives us information about the class [11].

$$IG(t) = -\sum_{i=1}^{n} p(c_i) \log p(c_i) p(t) \sum_{i}^{n} p(c_i|t) \log p(Ci|t) + p(ta) \sum_{i}^{n} p(c_i|ta)\, log\, p(Ci|ta) \qquad (4)$$

Where

C_i represents the i th class, $p(c_i)$ is the probability of the I the class.

P(t) and p(ta) are the probability that term t appears or not.

$P(c_i|t)$ is the conditional probability of the I th class given that term t appears.

$P(c_i|ta)$ is the conditional probability of the I th class given that term t not appear.

C) Classification

In this research two supervised machine learning algorithms RBF and FRBF have been used for TC. Based on the obtained feature from section 3.2 a classifier is created. As any classification model, there are two phases in the

classifier: training phase and testing. In the training phase a set of documents with known labels are introduced to the model for classification. Each document in this set is supposed to fit a predefined class. In the testing phase a set of new documents with unknown labels are introduced to the classifier model to specify a class for these documents. Two important algorithms for classification are used RBF and FRBF.

1- Radial Basis Function (RBF)

RBF is an artificial neural network that consists of many layers, input layer, hidden layer and output layer. The network is fully connected where each node in the input layer connected to every node in the hidden layer. The output of hidden layer is summed to produce the output. The hidden layer consists of Gaussian function., Which is the most famous and important of all statistical distributions [14]. RBF consists of two phases training phase and testing phase algorithm 1 shows the main steps of RBF for TC.

Algorithm: TC using RBF training phase
Input: set of document collection D, vector C set of desired classes
Output: vector C set of desired classes
Step1: compute features vector F (F1, F2, F3, F4) as explained in section 3.2 Step2: feed the F feature to the input layer Step3: For each class, take its center and assign to it a Gaussian function as follows $G(X) = \exp\left(\frac{\|x-u_j\|^2}{2\partial j^2}\right)$ Step4: Take the outputs of the class K and connect them directly to a max neuron.

The testing phase consists of feeding the document features to the RBF network, which produce the desired class label.

2- Fuzzy Radial Basis Function (FRBF)

FRBF network is intended by combining the principles of RBF with the fuzzy c-means algorithm. The main modification of FRBF is made at hidden layer. FRBF also consists of three layers the input layer corresponds to the input features while the output layer corresponds to desired classes, whereas the hidden layer represents the number of classes. The main idea of fuzzy c-mean based on allowing the object belong to more than one class with some weight. The sum of all weights must equal to one, when classes are well separated, a crisp classification of objects into class makes sense [15].

IV. Dataset and Evaluation Measures

In this model Reuters 21578 used which is a benchmark dataset for text classification [16]. The dataset divided into 22 files. The evaluation measures of classification mainly based on precision, recall and f-measure, they can define as follows [17].

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$F-measure \frac{2*precision*recall}{Precision+recall} \quad (7)$$

V. Experimental Results

As described in section 4 Reuters 21578 were used. Only seven topics were selected with 200 files for each topic, where 120 files for each topic used in the training phase and the remaining 80 files used in the testing phase. Table 1 shows the results of the proposed model for both RBF and FRBF.

Table (1): The results of the proposed model

	RBF	FRBF
Precision	0.85	0.89
Recall	0.87	0.91
F-Measure	0.86	0.90

The results showed the preference of FRBF over RBF. As known the most important problem of the TC is the misclassification between classes. This problem occurs because the overlap of text features for one class with another class. Based on the idea of FRBF that allowed the object to be assigned to more than one class, then choosing the most suitable class based on fuzzy theory, therefore the results of the FRBF are more accurate than RBF.

VI. Conclusions and Future Work

In this paper, TC method propose based on RBF and FRB, where set of features are extracted and both algorithms are used for TC purpose. The results showed the effectiveness of FRBF over RBF due to the using of fuzzy theory for solving the problem of misclassification that occur in most classification algorithms. Our future work includes studying more feature and computing feature weights to improve the performance of the FRBF algorithm.

References

[1] A Joulin,. E. Grave,.P. Bojanowski,; and T. Mikolov" Bag of tricks for efficient text classification." In EACL, 427–431. Association for Computationa Linguistics,2017.

[2] X.Zhang, and B.Wu" Short Text Classification Based on Feature Extension Using The NGram Model", International Conference on Fuzzy Systems and Knowledge Discovery (FSKD).2015.
[3] U.Gulden "Experimental evaluation of feature selection methods for text classification,", 9th International Conference on Fuzzy Systems and Knowledge Discovery, 2012.
[4] B.Harish and M. Revanasiddappa;"A Comprehensive Survey on various Feature Selection Methods to Categorize Text Documents", International Journal of Computer Applications (0975 - 8887) Vol. 164 - No.8, April 2017.
[5] M.Mowafy, A.Rezk, ,H.El-bakry "An Efficient Classification Model for Unstructured Text Document", American Journal of Computer Science and Information Technology,Vol 6, No.1 2018
[6] A.Conneau, S. Holger, B. Loïc,and L. Yann "Very Deep Convolutional Networks for Text Classification",. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics.2017, 1, 107–1116.
[7] R.Jodha,.G. Sanjay, K. Chowdhary. And A. Mishra, "Text Classification using KNN with different Features Selection Methods". International Journal of Research Publications(IJRP). 2018, 8, 1.
[8] L. Yao, C. Mao and Y. Lua" Graph Convolutional Networks for Text Classification Graph Convolutional Networks for Text Classification.", Computing Research Repository (CoRR),2019.
[9] H.Oufaida, O. Nouali,and P. Blache" Minimum redundancy and maximum relevance for single and multi-document Arabic text summarization.", Journal of King Saud University – Computer and Information Sciences,2014.
[10] Porter Stemming Algorithm: http://www.tartarus.org/martin/PorterStemmer
[11] M,Ramya. and P.Alwin , "Different Type of Feature Selection for Text Classification,",International Journal of Computer Trends and Technology (IJCTT) – vol. 10 No.2,2014.
[12] X.Yan , J. Gareth ,T.Jin, W. Bin and S, ChunMing ," A Study on Mutual information based Feature Selection for Text Categorization, "Journal of Computational information Systems 2007.
[13] A.John." MULTI-DOCUMENT SUMMARIZATION SYSTEM: USING FUZZY LOGIC AND GENETIC ALGORITHM", International Journal of Advanced Research in Engineering and Technology (IJARET), Vol.7, No. 1, PP. 30-40.2016.
[14] C.Dash, A,Behera, S.Dehur and S. Cho," Radial basis function neural networks: a topical state-of-the-art survey", Open Computer Science, Vol.6 No.1. doi:10.1515/comp-2016-0005 2016.
[15] S. Mitra and J. Basak ," FRBF: A Fuzzy Radial Basis Function Network.", Neural Computing & Applications, 10(3), 244–252.doi:10.1007/s521-001-8052-9 .2001
[16] D.David available at: http://www.daviddlewis.com/resources/testcollections/reuters21578
[17] Y.Wang, and L. Zhu. improved text classification method based on combined weighted model", Concurrency and Computation: Practice and Experience, e5140.doi:10.1002/cpe.5140,2019
[18] Maab Alaa Hussain,"A RADIAL BASIS NEURAL NETWORK CONTROLLER TO SOLVE CONGESTION IN WIRELESS SENSOR NETWORKS" Iraqi Journal for Computers and Informatics Vol. 44, 2018, pp. 44-48.

Dr. Zuhair Hussein Ali is presently working at AL-Mustansiriyah university college of Education department of computer science. Received his B. Sc from AL-Nahrain university college of science, department of computer science in1995, the M.Sc from Baghdad university college of science, department of computer science in 1999 and the Ph.D from Al-Technology University in 2018

Lec.Ameen abdulzahra Noor is presently working at AL-Mustansiriyah university college of Education department of computer science. Received his B. Sc from Al-Mustansiriyah college of Education department of computer science.in 2004,
the M.Sc from Iraqi Commission for computers & informatics in 2012.