ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2021.0325

IMPROVED FEATURE EXTRACTION ON TEXT DOCUMENTS USING NEURAL NETWORK MODEL

V. Kumaresan and R. Nagarajan
Department of Computer and Information Science, Annamalai University, India

Abstract

In natural language processing, the text clustering plays a major role on reducing the text dimensionality. However, the lack of data models has made the clustering algorithm to face sparsity problems. The integration with deep learning has resolved the problem of scarce knowledge on text documents. However, deeper architectures learn such redundant features, which limit the efficiency of solutions. In this paper, a complete extraction of features from text document using neural network model. The neural network model utilizes feed forward mechanism and a type of unsupervised learning that denoises the corrupted input features. The reconstructed feature is used for initialing the feed forward network. This method reduces the manual labelling in the process of screening. For evaluation, series of experiments are conducted to investigate the performance of the method over the text datasets with various conventional algorithms.

Keywords:

Text Document, Feature Extraction, Neural Network, Denoising

1. INTRODUCTION

The text clustering is widely used in many applications, such as information filters, recommendations, sentiment surveys, opinion mining and web searches [1]. Text clustering techniques are typically divided into two types of statistical and rule-based methods [2]. Statistics use mathematical model, while rules-based techniques require broad domains to establish rules that can classify specimens into a predetermined set of categories. Regulatory-based methods are not always implemented because it is difficult to create consistent laws, which do not have to be checked periodically.

The automatic clustering documents for text were used by researchers who used computer clustering [3] - [5] previously. In recent years, a number of algorithms have been proposed in order to significantly enhance the efficiency of the text clustering [6] – [9]. While functional selecting techniques to a certain extent reduce data dimension, traditional text clustering methods still discuss feature representation as banal algorithms using word-pack models that treat features as unigrams, n-gram or specific patterns [10]. Although this approach reduces the data dimension to a certain degree. Therefore the complete data history is not documented and the data sparse issue is faced.

In addition, the problem of data slimming is addressed through word embedding, in addition to syntactic and semi-anticipated knowledge of text data. Besides the efficient collection of contextual data, in-depth learning methodologies in texts clustering resolve data sparsity problems and are superior to state-of-the-art machine learning approaches [11] [12].

In order to extract a greater clustering of the features, researchers have mainly tried to construct deeper neural network architectures in computer-vision and NLP [13] [14]. In the case of small data sets, however, noisy architectures resulting from sampling noise are not only computer-complicated, but also relationships learned from neural network architecture [15]-[17].

In this paper, the study propose a thorough extraction of features from the text document using neural network model. The neural network model utilizes feed forward mechanism and a type of unsupervised learning that denoises the corrupted input features. The reconstructed feature is used for initialing the feed forward network. This method reduces the manual labelling in the process of screening. For evaluation, the study conducts a series of experiments to investigate the performance of the method over the text datasets. Experimental validation shows reduced the item required for screening where the trade-off with sensitivity is not compromised.

The outline of the paper is presented below: section 2 provides the related works, section 3 discusses the proposed model for feature extraction. Section 4 evaluates the work and section 5 concludes the entire model.

2. RELATED WORKS

This section provides an overview of the state-of-the-art selection algorithms based on statistical texts in text clustering approaches. Furthermore, recent methodologies are briefly listed for the deep learning graduation. The selection of features is considered an important function in clustering because they exclude corpus properties that are obsolete and irrelevant. In recent years, researchers have proposed various filter-based selection methods for improving document clustering efficiency [18].

The frequency of the document [19] is the simplest measure of ratings in training results, using a particular attribute in positive or negative class documents. Another simple feature selection algorithm is proposed by Lai et al. [20], which is a replicated neural convolutional system [21] two-way text clustering structure. This recurring system gathers contexts and learns word representations and creates less noise than the trivial window-centered network.

Aziguli et al. [2] has applied hybrid and deep methods to minimised noise and enhanced performance of feature extraction. Similarly, Jiang et al. [11] suggested to classify the text in order to solve the computation problem in the form of a sparse sparse matrix for the text clustering task, using a text clustering model for extraction.

Edinburgh et al. [22] proposed a hybrid deep-belief algorithm for the sentiment clustering. They first extracted the characteristics in their two fold networks using the convolutional RBM from previously secret layers with Boltzmann machines.

Huang et al. [23] used deep networks of faith to learn about emotional characteristics from speech signals. The clustering of

the non-linear SVM has been fed extracted characteristics to establish a hybrid emotional detection process.

Kahuet al. [24] has shown that the decay of the Relu units is probable, rather than max out units, to be further enhanced. Liu et al. [25] has a cautious framework included with knowledge of the meaning of corpus terms in neural network clustering.

3. PROPOSED NEURAL NETWORK MODEL ON TEXT FEATURE EXTRACTION

The proposed neural network automated feature extraction architecture is shown in Fig.1. The test takes place in the same experimental configuration and is divided into two sets of unlabeled and randomly labelled texts. The two sets consist of 50 percent of the text of the data set, while the unlabelling text of the label set does not overlap the text of the dataset. An integration/out of reach human analysis manually determines the marked array and makes a distinction between qualifying and unacceptable studies using the text clustering system. The study is mindful in experiments of using manually noted public data sets.

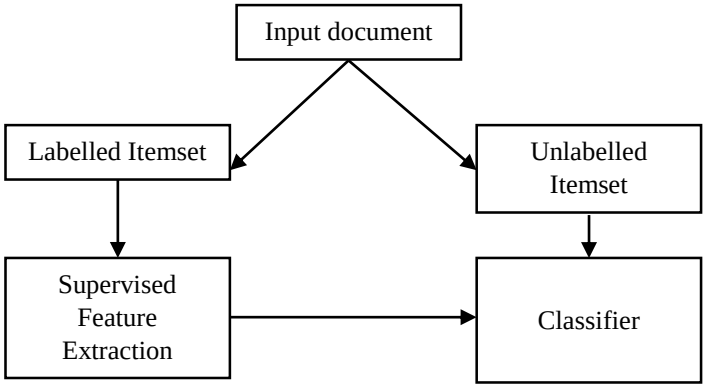

Fig.1. Automated Feature Extraction Architecture

In the text clustering procedure, a textual content extraction feature is first used for transforming text data into a numerated image, namely vectors of functionality. The study designs a new learning model with neural networks in order to extract text in our approach. This document is used as a input for a linear SVM clustering as the document extracted from our method.

The linear SVM clustering has been trained to distinguish between eligibility and inacceptable texts in the light of the above features. Clustering of linear SVM. A linear SVM produces a linear hyperplane that makes it possible to differentiate between credentials from inappropriate texts in data sets. Following a linear trend course in SVM, the study prioritise unreported text in the data set using the trained models, so that text is likely to be included in the study in higher rankings than text in less-classified datasets. In particular the text of the unlabeled dataset is defined by the confidence of class qualifying in the text of a data sequence. The confidence depends on the distance between the vector and the SVM hyperplane, i.e. the longer the distance, the higher the rating confidence. When the text in the class eligible dataset is prioritised, text in the highest data set is added to the survey, while the lowest-ranking text in the data set is deemed as inadmissible and is thus automatically excluded from the analysis. For a callback standard the cut-off point is 95%.

3.1 FEATURE EXTRACTION

The denoise autocoder is intended to restructure the BoW input space provided that the BoW version is corrupted. In a particular data set, the raw frequency is the value of the dimension of a term. Previous studies have shown how a denoise autoencoder learns how to extract input sound from conventional self-encoders that are trained in the input of cleaned intruding data. On this basis, the study corrupt the BoW input function using the Gaussian noise standard deviation additive.

The study use the simple DAE variation to get more complex non-linear input data projections, namely the deep DAE which adds more hidden, intermediate network layers. In addition, theoretically, three different DAEs are used to learn about different BoW reconstruction areas. Each DAE is composed of five cache layers, while the dimensions of the first and last BoW layers vary from one DAE to the other. The recomposed output of each DAE is then used to initialise the supervised supply of the neural network. This form of unregulated, pre-trained training has shown that the performance of the feed forward network improves significantly when the feed forward neural network is initialised through deep DAEs.

The neural feed-in network consists of 6 completely secret strata, i.e. $\{L_1, L_2, L_6\}$, and the softmax output stratum L7, which calculates the probability distribution over the permissible and inadmissible class for each text dataset. The first three layers of the $\{L_1, L_2$ and $L_3\}$ network are parallel, which means that there is no relation between the three layer units and they are initialised by the output reconstruction of the three DAEs. The three layers are then linked with a broad L_4 which is fully connected.

The neural feed forward system is monitored by reducing the interplay between the probability distribution of the gold groups and the possible class distribution of the softmax layer. The weights of the feed forward network shall be set during training with the vanilla stochastic gradient. The study extract functional vector systems that fit the entire set of data acquired after training in the feed forward network using the mass pattern of the widely connected L_4 layer.

The extracted document vectors, i.e. the output of the L_4 linked layer of the proposed SVM text clustering, are then used as data. It should be noted that the extraction stage does not rely on the clustering model and that the extraction process is not the clustering step of the document. As such, different methods of extracting functionality can be used for the same text clustering, and different text classifiers can have the same extraction process. In the course of this study, the effects of our existing extraction methods will be checked and evaluated. Therefore, the study compares the different basic extraction methods with our proposed process using the same linear SVM text classifier.

4. PERFORMANCE EVALUATION

This section describes the experimental setup used to evaluate the integrity of proposed text clustering methodology on a benchmark dataset namely BBC-Sports dataset [26].The dataset has 2225 documents collected from the BBC news website in correspondence with five topical areas between 2004 and 2005. The class labels includes athletics, cricket, football, rugby and tennis. The datasets have been pre-processed as follows:

stemming (Porter algorithm), stop-word removal (stop word list) and low term frequency filtering (count < 3) have already been applied to the data.

The performance is estimated in terms of accuracy, sensitivity, specificity, f-measure, percentage error and geometric mean represented between Fig.2 – Fig.7.

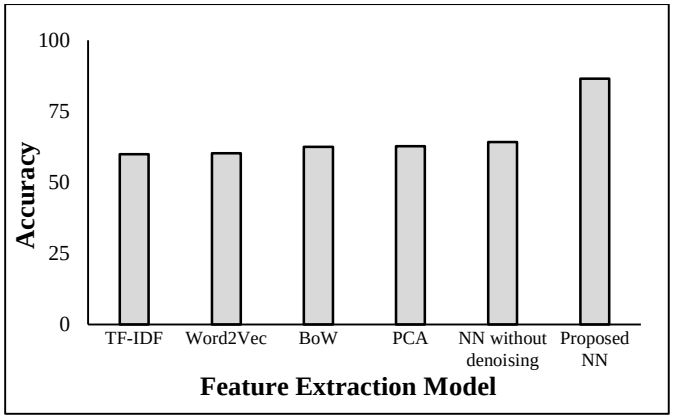

Fig.2. Accuracy

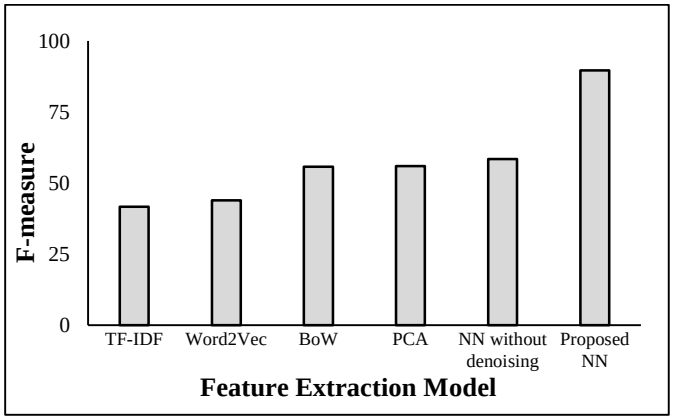

Fig.3. F-measure

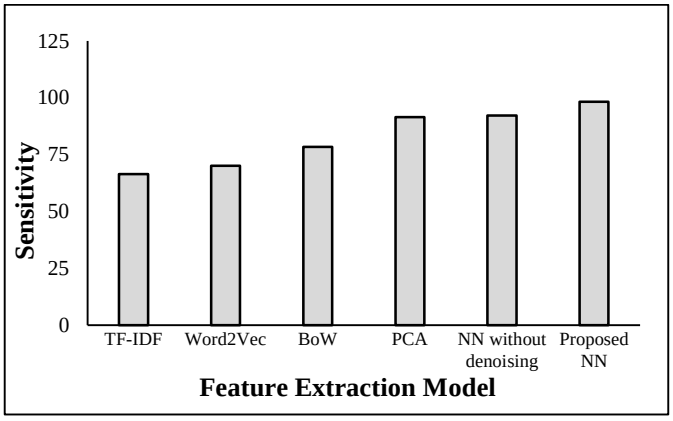

Fig.4. Sensitivity

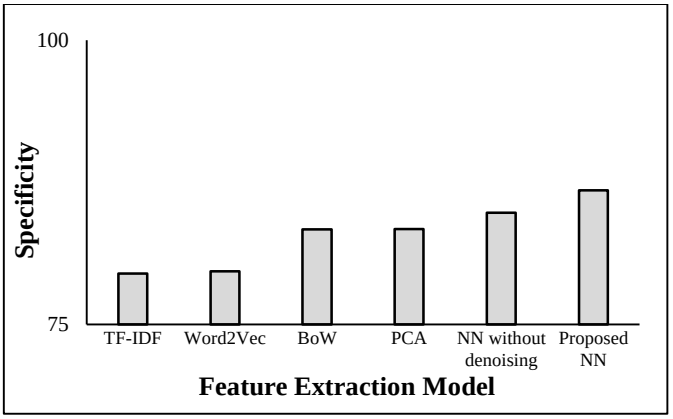

Fig.5. Specificity

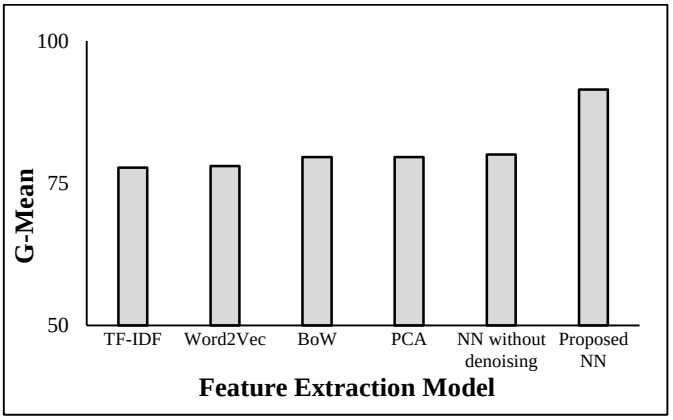

Fig.6. Geometric mean

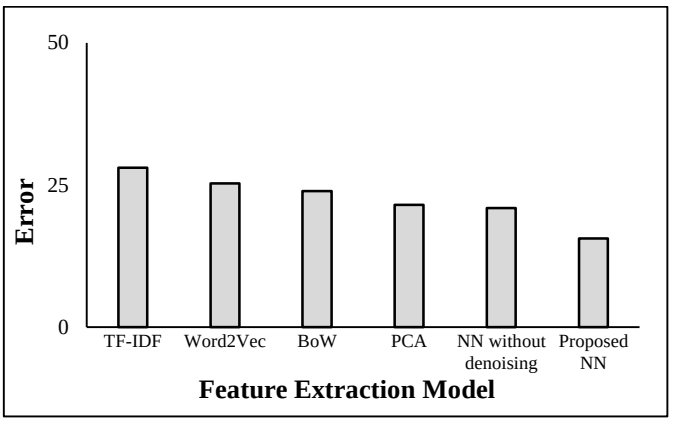

Fig.7. Percentage error

From the results of Fig.2 – Fig.7, the results shows that the proposed neural network model offers improved accuracy, f-measure, sensitivity, specificity, geometric mean and reduced percentage error than existing text feature extraction methods.

5. CONCLUSION

In this paper, the study focus on the optimal extraction of features from the text document using neural network model. The neural network model uses unsupervised learning mechanism to select the optimal features from the pre-processed text document. The selected features is used to configure the neural network. This

technique eliminates the screening process for manual labelling. The evaluation shows an improved system efficiency on text datasets throughout a series of experiments. The results shows improved accuracy and precision towards text data extraction than other methods.

REFERENCES

[1] C.C. Aggarwal and C. Zhai, "*A Survey of Text Classification Algorithms*", Springer, 2012.

[2] W. Aziguli, Y. Zhang, Y. Xie and D. Zhang, "A Robust Text Classifier based on Denoising Deep Neural Network in the Analysis of Big Data", *Scientific Programming*, Vol. 2017, pp. 1-20, 2017.

[3] L.E. Peterson, "K-Nearest Neighbor", *Scholarpedia*, Vol. 4, No. 2, pp. 1883-1887, 2009.

[4] P. Langley, W. Iba and K. Thompson, "An Analysis of Bayesian Classifiers", *Aaai*, Vol. 90, pp. 223-228, 1992.

[5] X. Luo, J. Deng, J. Liu and W. Wang, "A Quantized Kernel Least Mean Square Scheme with Entropy-Guided Learning for Intelligent Data Analysis", *China Communications*, Vol. 14, No. 7, pp. 1-10, 2017.

[6] T.N. Lal, O. Chapelle and J. Weston, "*Embedded Methods*", Springer, 2006.

[7] A. Rehman, K. Javed and H.A. Babri, "Feature Selection based on a Normalized Difference Measure for Text Classification", *Information Processing and Management*, Vol. 53, No. 2, pp. 473-489, 2017.

[8] R. Wald, T. Khoshgoftaar and A. Napolitano, "Filter-and Wrapper-based Feature Selection for Predicting user Interaction with Twitter Bots", *Proceedings of IEEE International Conference on Information Reuse and Integration*, pp. 416-423, 2013.

[9] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection", *Journal of Machine Learning Research*, Vol. 3, No. 2, pp. 1157-1182, 2003.

[10] Q. Le and T. Mikolov, "Distributed Representations of Sentences and Documents", *Proceedings of International Conference on Machine Learning*, pp. 1188-1196, 2014.

[11] M. Jiang, Y. Liang and X. Feng, "Text Classification based on Deep Belief Network and Softmax Regression", *Neural Computing and Applications*, Vol. 29, No. 1, pp. 61-70, 2018.

[12] C.H. Shih, B.C. Yan and S.H. Liu, "Investigating Siamese LSTM Networks for Text Categorization", *Proceedings of Asia-Pacific Conference on Signal and Information Processing Association Annual Summit*, pp. 641-646, 2017.

[13] C.Y. Lee, S. Xie, P. Gallagher and Z. Zhang, "Deeply-Supervised Nets", *Proceedings of International Conference on Artificial Intelligence and Statistics*, pp. 562-570, 2015.

[14] C. Szegedy, W. Liu, Y. Jia and P. Sermanet, "Going Deeper with Convolutions", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-9, 2015.

[15] M. Denil, B. Shakibi, L. Dinh and M.A. Ranzato, "Predicting Parameters in Deep Learning", *Proceedings of International Conference on Advances in Neural Information Processing Systems*, pp. 2148-2156, 2013.

[16] B.O. Ayinde, T. Inanc and J.M. Zurada, "On Correlation of Features Extracted by Deep Neural Networks", *Proceedings of International Conference on Neural Networks*, pp. 1-8, 2019.

[17] B.O. Ayinde and J.M. Zurada, "Clustering of Receptive Fields in Autoencoders", *Proceedings of International Conference on Neural Networks*, pp. 1310-1317, 2016.

[18] A. Rehman, K. Javed, H.A. Babri and M.N. Asim, "Selection of the Most Relevant Terms based on a Max-Min Ratio Metric for Text Classification", *Expert Systems with Applications*, Vol. 114, No. 1. pp. 78-96, 2018.

[19] A. Dasgupta, P. Drineas, B. Harb and V. Josifovski, "Feature Selection Methods for Text Classification", *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 230-239, 2007.

[20] S. Lai, L. Xu, K. Liu and J. Zhao, "Recurrent Convolutional Neural Networks for Text Classification", *Proceedings of International Conference on Artificial Intelligence*, pp. 1-14, 2015.

[21] N. Kousik, S. Kallam, R. Patan and A.H. Gandomi, "Improved Salient Object Detection using Hybrid Convolution Recurrent Neural Network", *Expert Systems with Applications*, Vol. 166, pp 1-20, 2020.

[22] S. Zhou, Q. Chen and X. Wang, "Active Semi-Supervised Learning Method with Hybrid Deep Belief Networks", *PloS One*, Vol. 9, No. 9, pp. 1-9, 2014.

[23] C. Huang, W. Gong, W. Fu and D. Feng, "A Research of Speech Emotion Recognition based on Deep Belief Network and SVM", *Mathematical Problems in Engineering*, Vol. 12, No. 3, pp. 1-16, 2014.

[24] S.E. Kahou, C. Pal, X. Bouthillier and P. Froumenty, "Combining Modality Specific Deep Neural Networks for Emotion Recognition in Video", *Proceedings of ACM on International Conference on Multimodal Interaction*, pp. 543-550, 2013.

[25] M. Liu, G. Haffari, W. Buntine and M. Ananda-Rajah, "Leveraging Linguistic Resources for Improving Neural Text Classification", *Proceedings of the Australasian Language Technology Association Workshop*, pp. 34-42, 2017.

[26] BBC Sports, Available at: http://mlg.ucd.ie/datasets/bbc.html.