2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT 2018)

Bengaluru, India
10-12 July 2018

Pages 1-789

IEEE Catalog Number: CFP1852J-POD
ISBN: 978-1-5386-4431-7

****** This is a print representation of what appears in the IEEE Digital Library. Some format issues inherent in the e-media version may also appear in this print version.***

IEEE Catalog Number:	CFP1852J-POD
ISBN (Print-On-Demand):	978-1-5386-4431-7
ISBN (Online):	978-1-5386-4430-0

Additional Copies of This Publication Are Available From:

Curran Associates, Inc
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: (845) 758-0400
Fax: (845) 758-2633
E-mail: curran@proceedings.com
Web: www.proceedings.com

TABLE OF CONTENTS

Domain Classification of Textual Conversation Using Machine Learning Approach

Sandeep Rathor[1], R. S. Jadon[2]

[1]Assistant Professor, Dept of CEA, GLA University, Mathura, Uttar Pradesh, India.

[2]Professor, Dept of MCA, MITS, Gwalior, Madhya Pradesh, India

Abstract- This paper presents an approach for classification of textual conversation into multiple domain categories using support vector classifier. The feature reduction is done through Principal Component Analysis (PCA) to extract the important features from the feature vector. These features are passed to different configurations of SVM and the best one is chosen for the final process of classification. The domain's categories are defined on real life situations and conversation to train the system like education & research, personal, patriotism, terrorism, medical, religious, sports, and business. The experiment results show that the proposed method works effectively with more than 75% accuracy.

Keywords- Machine Learning, Support Vector Classifier, Domain Classification, Textual Conversation, Machine Intelligence.

I. INTRODUCTION

The text is the basic means of communication. It not only communicates informative contents, but also produces some additional information like feelings, emotions, sentiments and domain. The Domain recognition is different to sentiment analysis and emotion recognition. The objective of sentiment analysis is to sense positive, neutral, or negative feelings from text, whereas emotion plays an elementary role in our daily lives. Its objective is to detect and recognize the feelings through the text, such as anger, fear, disgust, happy, sad, and surprise[1]. Generally Ekman's six emotion class is used to detect the emotions. The Sentiment analysis can be classified into two categories; opinion mining and emotion mining. Opinion mining concern with the expression of opinions, for example neutral, positive or negative while emotion mining concerned with the pronunciation of emotions like sad, happy, excited etc.[2].

Domain recognition is one of the fields of affective computing. It refers to evaluating conversation towards different issues. If two persons are discussing their personal problem through text conversation then the category of domain is personal therefore; our system is capable enough to recognize its domain categories automatically for which the conversation has taken place while emotion recognition or sentiment analysis only reflect the mood or emotions. Classifying domain from the textual conversation can be applied to the various applications, based on human computer interaction. It is a relatively new classification and advancement in the field of affective computing and machine learning. It can also be useful to prevent miss happening or unwanted activities on the basis of text conversation. We can recognize different types of domains on the basis of communications or conversation. The categories of domain are defined on the basis of various real life situations and conversation. Every conversation or statement always may fall

into any one of the domain category, such as; education & research, personal, patriotism, terrorism, medical, religious, sports, and business. A large paragraph may provide the largest number of clues as features for domain analysis, however, least number of features reflect feature sparseness problem [3].

The aim of the proposed approach is to classify the domain of communication or conversation through textual contents with acceptable accuracy.

The proposed paper consists of five sections as follows: Section 2 related work proposed by various researchers section 3 contains the proposed work and Section 4 represents the result of the proposed scheme, finally, at last section 5 represents the conclusion and future scope of the proposed scheme.

II. RELATED WORK

It is a novel approach, therefore, no work is done in the same context, however, the similar approach is used by various researchers on sentiments and emotion recognition as:

[4] Proposed the application of spoken language understanding between human and machine through the deep belief network. Text classification algorithms used in this paper are SVM, boosting and maximum entropy. However, Result analysis in this paper shows that SVM effectively perform non-linear classification to high dimensional feature space and flexibility of changing the choice of the kernel. The working of linear kernel is much faster in comparison to other kernel. The approach used for text classification is very nice, however, overall process requires more computation time due to create deep network.

Mathematical description and model of different machine learning techniques like supervised, unsupervised, semi-supervised along with its impact is discussed [5]. The essential intention of learning is to find a decision function that is able to predict the output of future. In general, the prediction task is called classification when the output takes categorical values. In this paper, author also analyzed recent developments of deep learning and learning with sparse representations, focusing on their direct significance.

[6]Proposed an approach for emotion recognition from text. In this paper emotion generation rules are manually defined and semantic labels are used to represent an emotion state through the emotion association rules. A separable mixture model is used to estimate the similarity between an input sentence and emotion association rules of each emotional state. There are only three emotional states used in this paper for performance evaluation, i.e. happy, unhappy, and neutral, however, in real situation emotion state may vary from the defined states. Therefore, it has very limited emotion states.

[7] Proposed a multimodal approach for emotion recognition from speech and text using support vector machine. In this paper emotional keywords defined manually while emotion modification words assigned an integer value (positive or negative). A higher value represents more impact on emotion. For example, emotion modification value for the word, ‘good’, ‘very good’ and ‘extremely good’ are +1,+2,+3 respectively. Therefore extremely good have the highest value. If there is negative sentence, then these values are assigned as negative. They also discussed about neutral state of emotion, if emotion intensity value is lower than predefined threshold, however, other states of emotion is not defined in this paper. The training set used in

this paper is also limited. [2] Proposed an analysis of the current state of text sentiment, one step ahead on sentiment analysis. There are two concepts used in this paper; Opinion mining and Emotion mining. Opinion mining the reflect attitude by using natural language processing. Emotion mining reflects emotions regarding topics by using machine learning technique. Parts Of Speech tagging is used to determine the sense of the text. This paper focused on the sentiment analysis on different way only. However, there is no way to find the domain of text paragraph discussed in this paper.

Detection and classification of acoustic scenes and event is proposed by [8] on the basis of audio signals. However, in an environment, events are overlapping or polyphonic whereas an audio signal is monophonic in nature. Therefore, to focus on a single event noise is spatially or spectrally suppressed in order to focus on one source of the event. On the basis of audio signal author proposed a method to recognize the environment like bus, office, street, park, restaurant, etc. and also discussed the classification of specific events presented by various researchers from the environmental audio like bird song, musical instruments and other harmonic sounds. Discussed events in this paper were very limited and due to the polyphonic environment the recognition of scenes and event is less accurate. Proposed detection and analysis of emotion from human speech is proposed in [9] . MFCC is used to classify the speech signal. From the speech signals we cannot find good accuracy because many of the human do not change their speech characteristics (like pitch, frequency, entropy etc.) at the time of the conversation. Therefore, the system is not able to find the emotion for this type of situation. The emotion used in this paper was also limited.

Works proposed by the researchers are focused on either emotion or sentiments While no one has discussed about domain classification and recognition. The objective of the proposed method is to classify the domain of communication through textual contents with acceptable accuracy.

III. PROPOSED WORK

A proposed framework for domain classification includes five phases; Lexical Analysis, Features Extraction, Features Scaling, Features Reduction and Domain Classification & Recognition as shown in the figure 1. Useful words are extracted using lexical analysis from the input text paragraph or sentences. Then these words transform into a mathematical form for further processing through feature extraction and normalize these values because it is required to apply machine learning algorithm. After normalization, try to reduce the features through PCA. PCA is an unsupervised technique used for feature reduction in data analysis and machine learning. It transforms an N x M matrix of data set to N x K matrix were M<K. It actually maximizes the total variance of the independent variables. This is done by projecting the actual data set on a lower plane. Now, we need to classify the category of domain on the basis of classifiers.

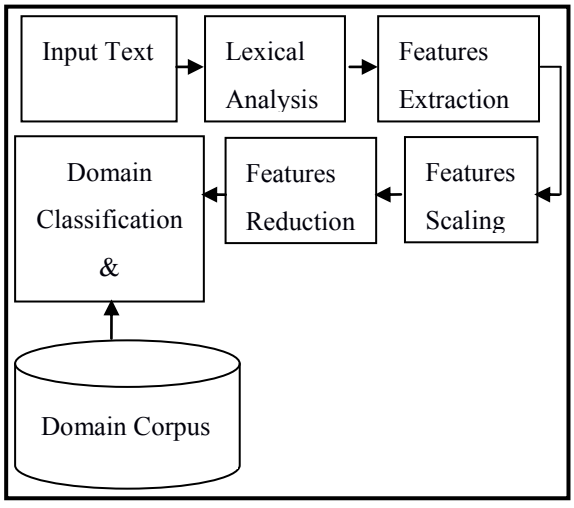

Figure 1. Proposed Framework for Domain Classification of Textual Conversation.

Every sentence or paragraph represents some clue about which conversation has taken place. In this context, the domain is classified into 8 categories, i.e., education & research, personal, patriotism, terrorism, medical, religious, sports, and business. The support vector machine has been extensively functioning in many research areas such as pattern matching, linear regression, data mining, data clustering etc. [7]. Therefore, to classify the text sentences into various domain categories, we used support vector machine with different configurations out of which select the best that produces the highest accuracy. The objective of SVM is to find a hyperplane that can completely distinguish different classes and it is decided by the maximal margin of classes. The samples that lie in the margin are called support vectors[10]. The equation of hyperplane is represented as an equation (1)

$$D(x) = w^T x + b \qquad (1)$$

where w is weight matrix and b is the intercept.

SVM can efficiently perform non-linear classification to transform the data using a technique called the kernel trick. Based on these transformations, it finds an optimal boundary between the possible outputs. Much of the flexibility and classification power of SVM depends on the choice of kernel[4]. Generally, the kernel used in SVM are linear, polynomial, radial basis function (RBF) and sigmoidal. The selection of kernel depends on the nature of the data used for classification. Every kernel has its own pros and cons. A linear kernel is applied when the data is linearly separable and uses linear functions. Generally, linear kernel is suitable for text classification because the text is often linearly separable but in our case it does not provide the best results. Polynomial kernel may lead to overfitting therefore; it may cause the problem of generalization. The selection of polynomial kernel is suitable when we have discrete data with no natural notion of smoothness. The rbf kernel uses normal curves around the data points, and sums these so that the decision boundary can be defined by a type of topology condition such as curves where the sum is above a value of 0.5[11]. RBF kernel is the most popular kernel because nonlinearly map samples into a higher dimensional space unlike to linear kernel and it also has less hyperparameters than the polynomial kernel. The sigmoidal kernel is suitable when the offset parameter is negative and close to 0 and problems where the number of dimensions is high or non linear separation in 2 dimensions.

IV. RESULT AND ANALYSIS

For the purpose of system evaluation, inorder to recognize the domain of text sentences, we collected 2713 sentences for training and 950 sentences for testing which includes 2520 different features. The domain of text sentences was tagged manually. The domain classification results are listed in the table 1. The table depicts various domains for classification and their respective number of sentences that are tagged. To evaluate the accuracy of the proposed system we used this tagged data as the target and passed the same sentences to the proposed system for prediction. The data set is comprised of sentences from various domains taken from standard sources of information like newspaper, social networking site, journal articles, etc.

Table 1. Tagged domain classification of textual sentences in the testing dataset.

Domain	No of Tagged Sentences
Religious	150
Research	90
Business	130
Sport	110
Medical	100
Terrorism	130

Patriotism	130
Personal	110

Experimental Results of PCA:

To train and test the performance of the proposed system we have collected total 3663 sentences. Out of which 2713 sentences are used for training and 950 sentences are used for testing which includes 2520 different features. After using PCA, 2520 features are reduced to 1495 only with threshold variance of 96%.

Variance represents a measurement of the variability in the data set. If we want a single principal component with the most variability, we need to search for maximum variance so that the principal component collects the most uniqueness from the data set. The graph represented in figure 2, shows the relationship between principal component and their respective variance where it is evident that the first principal component has the highest variance and remaining principal components exhibit decreasing variance.

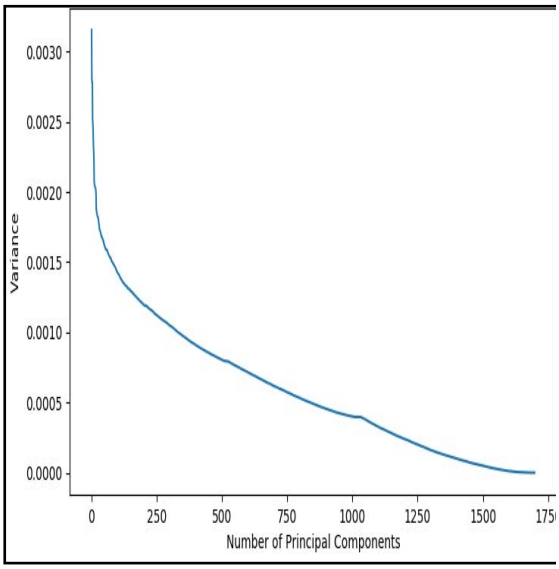

Figure 2. Relation between principal components and their variance.

Experimental Results of SVM:

In order to determine the correct SVM configuration we validate different SVM configurations on validation set. Validation set and training set used in proposed paper are divided in to ratio of 1:4.

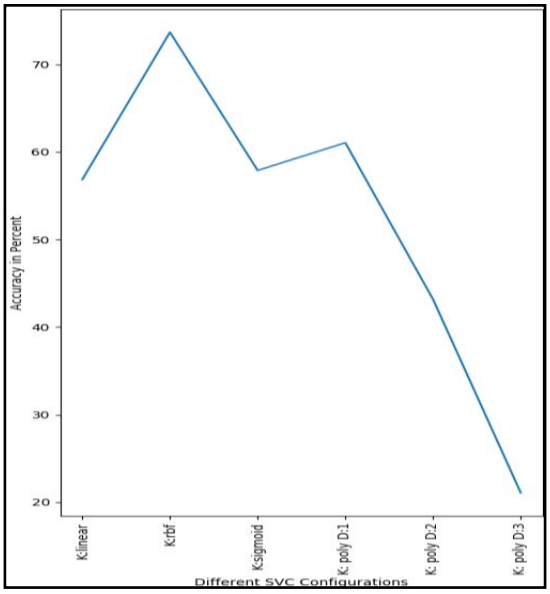

Figure 3. Cross Validation; Accuracy percentage of different SVM kernel configurations; kernel= linear, kernel=rbf , kernel=sigmoidal, kernel= polynomial with degree=1, 2 and 3.Relation between principal components and their variance.

Figure 3 shows the result of different SVM kernel configurations on validation set. Different configurations of kernel are taken as kernel= linear, rbf, sigmoidal and polynomial with the degree=1, 2 and 3. However, maximum accuracy is received when kernel selected as rbf as shown in figure 3. RBF kernel is one of the most popular kernel functions.

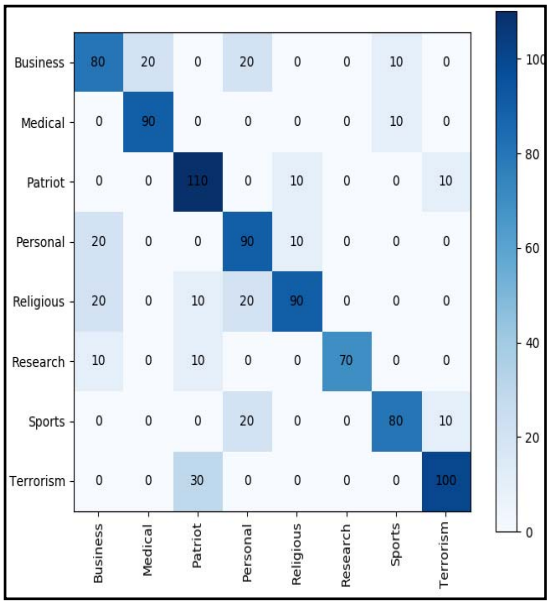

Figure 4. Confusion matrix of Domain Classification of Test Data.

Figure 4 is the confusion matrix that shows the correct and incorrect classification and recognition of text sentences based on a particular SVM configuration (kernel=rbf, C=1, γ=1) which is selected through a validation process as already shown in figure 3. The diagonal values represent the true positives or the correct classification of the domain, whereas the other values represent the mis classification of the domain. The figure 4 shows that the correct classification and recognition is more than 75%.

V. CONCLUSION AND FUTURE SCOPE

The paper explored the idea of domain classification for textual conversation. It was observed that domain recognition differs from emotion recognition and sentiment analysis because emotion recognition or sentiment analysis only reflects the mood or emotions while domain recognition recognizes its domain categories automatically for which the conversation has taken place. It is a relatively new classification and advancement in the field of machine learning.

VI. REFERENCES

[1] Che-Wei Huang and Shrikanth. S. Narayanan, "Characterizing Types of Convolution in Deep Convolutional Recurrent Neural Networks for Robust Speech Emotion Recognition," pp. 1-19, 2017.

[2] Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R. Zaiane, "Current State of Text Sentiment Analysis from Opinion to Emotion Mining," *ACM Computing Surveys*, 2017.

[3] Ahmed Abbasi, Hsinchun Chen, Sven Thoms, and Tianjun Fu, "Affect analysis of web forums and blogs using correlation ensembles," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 9, pp. 1168-1180, 2008.

[4] Ruhi Sarikaya, Geoffrey E. Hinton, and Anoop Deoras, "Application of deep belief networks for natural language understanding," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 4, pp. 778-784, 2014.

[5] Li Deng and Xiao Li, "Machine learning paradigms for speech recognition: An overview," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 5, pp. 1060-1089, 2013.

[6] Chung-Hsien Wu, Ze-Jing Chuang, and Yu-Chung Lin, "Emotion recognition from text using semantic labels and separable mixture models," *ACM Transactions on Asian Language Information Processing*, vol. 5, no. 2, pp. 165-183, 2006.

[7] Z J Chuang and Chung-hsien Wu, "Multi-modal emotion recognition from speech and text," *Journal of Computational Linguistics and Chinese*, vol. 9, no. 2, pp. 45-62, 2004.

[8] Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D. Plumbley, "Detection and Classification of Acoustic Scenes and Events," *IEEE Transactions on Multimedia*, 2015.

[9] Assel Davletcharova, Sherin Sugathan, Bibia Abraham, and Alex Pappachen James, "Detection and Analysis of Emotion from Speech Signals," *Procedia Computer Science*, vol. 58,

pp. 91-96, 2015.

[10] Simon Tong and Daphne Koller, "Support Vector Machine Active Learning with Applications to Text Classification," *Journal of Machine Learning Research*, pp. 45-66, 2001.

[11] E. A. Zanaty, "Support Vector Machines (SVMs) versus Multilayer Perception (MLP) in data classification," *Egyptian Informatics Journal*, vol. 13, no. 3, pp. 177-183, 2012.