IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-6, June 2018 E-ISSN: 2347-2693

Sentiments and Domain Analysis of Text Sentences Using POS Tagging & Machine Learning Approach

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

[1] Dept. of MCA MITS, RGPV Bhopal, India
[2] Dept. of MCA MITS, RGPV Bhopal, India

[*]*Corresponding Author: sandeeprathorresearch@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 17/Jun/2018, Published: 30/Jun/2018

***Abstract*—** This paper presents an efficient approach for sentiment and domain analysis of text sentences using POS tagging and random forest classifier machine learning approach. POS tagging technique is used for sentiment analysis while Random forest classifier is used for domain analysis of the text sentences. Various categories of sentiments are defined as positive, neutral, and negative while the domain's categories are defined on various real & professional life sentences to train the system like education & research, personal, marketing & advertizement, security of nation, political, religious, sports and legal issues. Every text sentence always reflects the domain's categories along with its sentiments. Therefore, Analyzing domain of text sentences along with sentiments is a challenging task and can be useful for various applications based on human computer interaction. The experimental result shows that the proposed method works effectively, efficiently and can be applied on real life applications where obligatory actions are taken automatically through sentences.

***Keywords*—** Random forest classifier, Machine Learning, Domain Analysis,Sentiment Analysis, Human computer interaction.

I. INTRODUCTION

Communication is a Latin word and it is the basic necessitate of life. The most common medium of communication is a language. Besides, there are other several means of communication available to us. We also use non-linguistic symbols such as traffic lights, road signs, railway signals to convey information relating to the movements of vehicles and trains however the most used forms in real life scenario are text, audio or video, facial expression etc. Through the audio we can recognize speech or speaker [1]. However, text is the most famous formal medium of communication. Every legal document is in the form of text and is also processed in the form of text from top to bottom or bottom to top. The basic means of communication is just transfer the information to others. However, more than that it includes domain for which communication has taken place as well as mood or sentiments also. A simple sentence may have the different sagacity. It depends on sentiments or mood of communicator. The objective of sentiment analysis is to sense positive, neutral, or negative feelings from the text sentences [2]. The Sentiment analysis can be classified as 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, exited etc.[3].

Domain analysis differs from emotion recognition or sentiment analysis. The emotion recognition or sentiment analysis only reflects the mood or emotions however, domain recognition recognizes domain categories automatically for which real or professional conversation. Domain's categories are defined on the various real life & professional life situations to train the system like education & research, personal, marketing & advertizement, security of nation, political, religious, sports and legal issues. A domain analysis approach along with sentiment analysis is more advantageous for current scenario. Moreover, it can also be implementing to the various real life applications, based on human computer interaction. A large paragraph may provide the larger number of clues as features for domain analysis, however, less number of features reflect feature sparseness problem [4].

The aim of the proposed approach is to classify the domain of text conversation with its sentiments through textual contents with acceptable accuracy in lower computation time.

Rest of the paper is organized as follows, Section I contains the introduction of of communication, domain analysis and sentiment analysis, Section II contain the related work in the field of sentiments analysis and POS tagging & keyword spotting, Section III contain the proposed methodology, Section IV contain the result analysis of proposed approach, Section V explain the research work with future direction.

II. RELATED WORK

There is no prior researchwork available in the field of domain analysis. However, various techniques of POS tagging & keyword spotting and machine learning for sentiment analysis proposed by various researchers are discussed in this section.

Keyword spotting deals with the identification of keywords in utterances. It is used to classify the sentiments in a sentence. The aim of Keyword Spotting is to detect predefined keywords in text sentences. On the basis of these predefined words we can classify the sentiments i.e. positive, neutral and negative. Sentiment analysis can be performed at different granularity levels, e.g., subjectivity detection simply classifies data as either subjective or objective [5], while polarity detection focuses on determining whether subjective data indicate positive or negative sentiment. Positive sentiment can be caused by happiness and negative sentiment can be caused by disgust while neutral is the case of between positive and negative. Keyword spotting can be applied to the audio video also. We cannot find good accuracy in speech signals because many of the human do not change their basic speech characteristics at the time of communication therefore, to process the text is good choice for domain recognition and sentiment analysis for accuracy point of view recurrent network approach may be good but time consuming [6]. Detection and classification of acoustic scenes and event is proposed by [7] on the basis of audio signals. However, in an environment, events are overlapping i.e. noise of multiple noise is mixed. Therefore, to focus on a single event noise is spatially or spectrally suppressed in order to focus on one source of the event. In this paper 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 while in our proposed paper domain of conversation along with sentiments can be find accurately. A multimodal approach for emotion recognition from speech and text using support vector machine is proposed by [8]. In this paper Author defined keywords manually to detect various emotion state and assigned an integer value positive or negative. A positive value have higher impact while a negative value have lesser impact of emotion. It is a simple type of sentiment analysis proposed by author. The integer value for a word, 'good teacher', 'very good teacher' and 'extremely good teachere' are +1,+2,+3 respectively. Therefore extremely good teacher has the highest value. Moreover, if there is negativity in sentence then values are assigned as negative.

III. PROPOSED METHODOLOGY

To the best of our knowledge, domain analsis is a new paradigm of research that is not yet widely recognized. A proposed framework for sentiment analysis and domain analysis can be represented as:

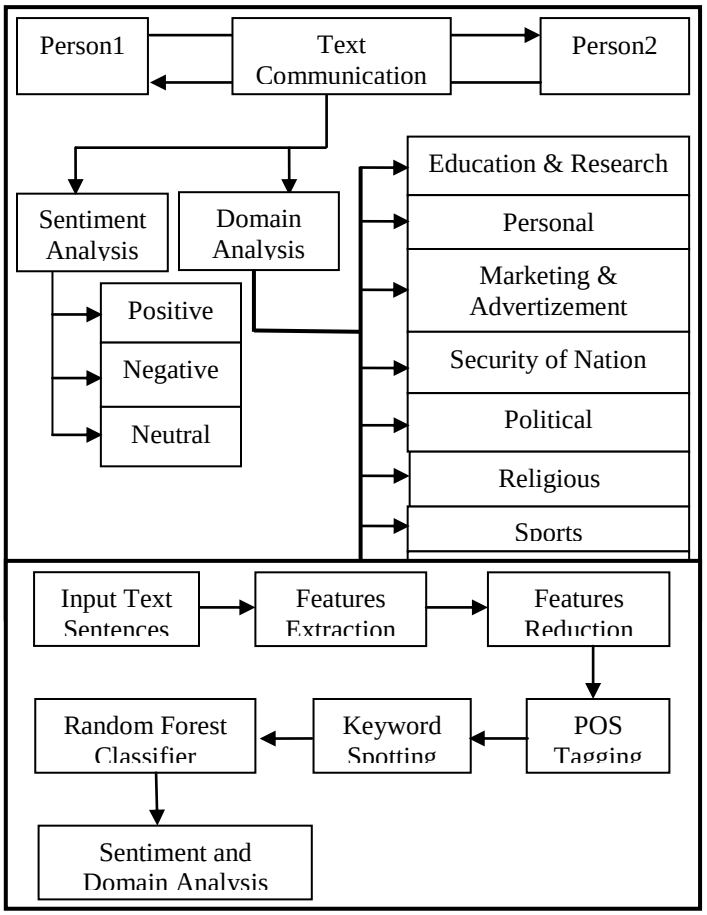

Figure 1. Proposed Framework for Sentiments and Domain Analysis of Text Sentences.

According to real environment sentiments can be positive, neutral and negative and the domain categories can be as education & research, personal, marketing & advertizement, security of nation, political, religious, sports and legal issues as shown in the figure 1. Every text sentence always may fall in to one of the defined domain with sentiment. Sentiment analysis indicates that how much positive or negative thought of a person in text conversation while neutral indicates between positive and negative.

A proposed framework for sentiments and domain analysis using POS tagging, keyword spotting and random forest classifier is shown in the figure 1. We applied the proposed method on text sentences because on speech signal we can not find as good accuracy as the text. Initially, the basic NLP operations are applied to the input text sentences that includes feature extraction and feature reduction along with POS tagging, and keyword spotting. Random forest classification is used for domain analysis purpose. Principal component analysis is done to identifying patterns in data, [9]. The result analysis shows that the performance of proposed system is satisfactory i.e. more than 75% and can be implemented in real life applications.

In the proposed scheme we used is Random forest classification because it combines more than one algorithms of same or different kind for classifying the objects [10]. For example, prediction is done through Naive Bayes, SVM and Decision Tree and then choose the best one for final consideration. Random forest classifier creates a set of decision trees from randomly selected subset of training set. It works with good accuracy because a single decision tree may be prone to a noise, but aggregate of many decision trees reduce the effect of noise and give more accurate results.

IV. RESULTS AND DISCUSSION

To test the performance of the proposed system, we collected 1500 sentences for training and 848 sentences for testing which includes 1930 different features. The domain and sentiments of text sentences was tagged manually. The result of sentiment and domain analysis are listed in the table 1. To evaluate the accuracy of the proposed system we used this tagged data as the target and passed the same to proposed system for prediction. The accuracy of the proposed system can be represented through the confusion matrix. Diagonal elements of the confustion matrix show the correct classification and rest values show the miss classification as shown in figure 2. Values from {0-7} represent the domain's categories as {education & research, personal, marketing & advertizement, security of nation, political, religious, sports and legal issues} respectively.

Table 1. Tagged Domain & Sentiment classification of text sentences in the testing dataset.

Domain's Categories	No of Tagged Sentences	*Sentiments Categories*		
		Positive	Neutral	Negative
Education & Research	96	35	35	26
Personal	122	52	35	35
Marketing & Advertizement	98	35	35	28
Security of Nation	103	50	43	10
Political	110	50	35	25
Religious	102	55	35	12
Sports	89	40	45	4
Legal Issues	128	50	60	18

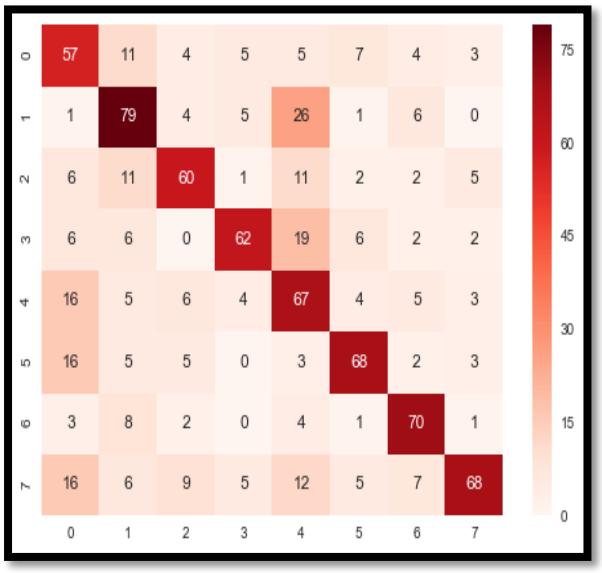

Figure 2. Confusion Matrix using Random Forest Classifier for Sentiment & Domain Analysis of Text Sentences.

V. CONCLUSION AND FUTURE SCOPE

The sentiment and domain analysis scheme for text sentences using keyword spotting and random forest classifier machine learning approach is presented in this paper. It can have many applications in the present scenario. On the basis of sentences a system is capable enough to recognize or analyse its sentiments and the domain. It is also implement to prevent mishappening on the ground of text conversation. In future, we will work to improve the accuracy.

REFERENCES

[1] S. Rathor, R. S. Jadon, "*Text independent speaker recognition using wavelet cepstral coefficient and butter worth filter*", 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1-5, 2017. doi:10.1109/ICCCNT.2017.8204079.

[2] A. Yadollahi, A. G. Shahraki, and O. R. Zaiane, "*Current State of Text Sentiment Analysis from Opinion to Emotion Mining*," ACM Computing Surveys, 2017.

[3] A. Abbasi, H. Chen, S. Thoms, and T. 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] S Dahl, George E., Dong Yu, Li Deng, and Alex Acero. "*Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition*", IEEE Transactions on Audio, Speech and Language Processing , vol 20, no. 1, pp 30-42, 2012.

[5] I. Szoke, P, Schwarz, P.Matejka, L. Burget, M. Karafiat and J. Cernocky, "*Phoneme Based Acoustics Keyword Spotting in Informal Continuous Speech"* Speech and Dialogue. Springer, Berlin, Heidelberg vol 3658. pp. 302–309,

[6] V. Krakovna and F. Doshi-Velez, "*Increasing the Interpretability of Recurrent Neural Networks Using Hidden Markov Models",* ICML Workshop on Human Interpretability in Machine Learning (WHI 2016).

[7] D.Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, *"Detection and Classification of Acoustic Scenes and Events," IEEE Transactions on Multimedia, 2015.*

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

[9] D. N. Agrawal and D. Kapgate, "*Face Recognition Using PCA Technique*", International journal of Computer science and Engineering, Vol. 2, Issue 10, pp. 59-61, 2014.

[10] A. D Kulkarni and B. Lowe, "*Random Forest Algorithm for Land Cover Classification*", International Journal on Recent and Innovation Trends in Computing and Communication, Vol. 4, Issue 3, pp.58- 63,2016.

Authors Profile

Mr. S. Rathor received his M. Tech degree in Computer Science and Engineering from Uttar Pradesh Technical University, Lucknow, India. He has 14 years of teaching and research experience and currently, he is a research scholar in the Department of Computer Applications, MITS, RGPV, Bhopal, India. His research interests include image processing and pattern recognition, natural languange processing, theory of computation.

Prof. R. S. Jadon has received his PhD degree in Computer Science & Engg. from IIT Delhi. He has more than 25 years of teaching and research experience, and currently working as a professor & Head in Department of Computer Applications, MITS, RGPV Bhopal, India. His research interests include Image Processing & Multimedia systems , Machine Learning and signal processing.

