ADVANCED MACHINE LEARNING TECHNIQUE FOR EFFICIENT CLASSIFICATION OF TEXT DATA

Ms. Konduri kavya #1, Ms. Garikapati Jaswitha #2, Ms. Boina Aneesha #3, Ms. Chillapalli Chandini #4, Mr. Nannuri suresh #5

#1 Student, Dept Of CSE, Qis College of Engineering and Technology, Ongole, Prakasam (Dt)
#2 Student, Dept Of CSE, Qis College of Engineering and Technology, Ongole, Prakasam (Dt)
#3 Student, Dept Of CSE, Qis College of Engineering and Technology, Ongole, Prakasam (Dt)
#4 Student, Dept Of CSE, Qis College of Engineering and Technology, Ongole, Prakasam (Dt)
#5 Assistant professor in Dept Of CSE, Qis College of Engineering and Technology, Ongole, Prakasam

Abstract

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. This paper displays a methodology for arrangement of printed discussion into various space classifications utilizing bolster vector classifier. The element decrease is done through Principal Component Analysis (PCA) to extricate the essential highlights from the element vector. These highlights are passed to various designs of SVM and the best one is picked for the last procedure of characterization. The space's classes are characterized on genuine circumstances and discussion to prepare the framework like training and research, individual, nationalism, fear based oppression, restorative, religious, sports, and business. The investigation results demonstrate that the proposed strategy works adequately with over 75% precision.

Keywords: *Machine Learning, Support Vector Classifier, Principal component Analysis.*

I. INTRODUCTION

The content is the essential methods for correspondence. It communicates instructive substance, yet likewise creates some extra data like emotions, feelings, estimations and space. The Space acknowledgment is distinctive to estimation investigation and feeling acknowledgment. The goal of notion examination is to detect positive, impartial, or negative emotions from content, while feeling assumes a rudimentary job in our every day lives. Its goal is to identify and perceive the sentiments through the content, for example, outrage, dread, appall, upbeat, miserable, and surprise [1]. For the most part Ekman's six feeling class is utilized to recognize the feelings. The Sentiment investigation can be grouped into two classifications; assessment mining what's more, feeling mining. Supposition mining concern with the declaration of suppositions, for instance nonpartisan, positive or negative while feeling mining worried about the elocution of feelings like miserable, glad, energized etc.[2]. Space acknowledgment is one of the fields of emotional figuring. It alludes to assessing discussion towards various issues. In the event that two people are talking about their own concern through content discussion then the class of area is close to home along these lines; our framework is sufficiently competent to perceive its area classes consequently for which the discussion has occurred while feeling acknowledgment or estimation investigation just mirror the inclination or feelings. Grouping space from the literary discussion can be connected to the different applications, in view of human PC cooperation. It is a moderately new arrangement what's more, headway in the field of full of feeling figuring and AI. It can likewise be helpful to anticipate miss occurring or undesirable exercises based on content discussion. We can perceive distinctive sorts of areas on the premise of correspondences or discussion. The classes of space are characterized based on different genuine circumstances and discussion. Each discussion or articulation dependably may fall into any of the space class, for example, instruction and research, individual, energy, fear mongering, medicinal, religious, sports, and business. A substantial passage may give the biggest number of signs as highlights for area examination, in any case, least number of highlights reflect highlight meager condition issue [3]. The point of the proposed methodology is to characterize the space of correspondence or discussion through printed substance with satisfactory precision.

II RELATED WORK

[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.

III PROPOSED SYSTEM

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

IV METHODOLOGY

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.

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].

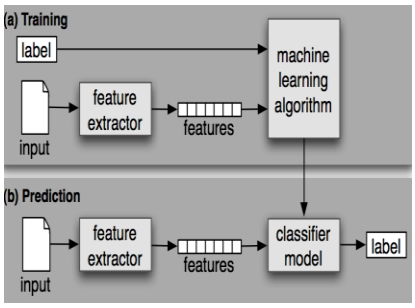

Fig: Framework of Proposed System

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 hyper parameters 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.

V CONCLUSION

The paper investigated the possibility of space characterization for literary discussion. It was seen that area acknowledgment varies from feeling acknowledgment and opinion investigation since feeling acknowledgment or supposition examination just mirrors the disposition or feelings while space acknowledgment perceives its area classes consequently for which the discussion has occurred. It is a moderately new order and progression in the field of AI.

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, IEEE – 43488 9th ICCCNT 2018 July 10-12, 2018, IISC, Bengaluru Bengaluru, India 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.

Authors Profile

Ms. **Konduri kavya** pursuing B Tech from computer science engineering in Qis college of Engineering and Technology(Autonomous & NAAC 'A' Grade), Pondure Road, vengamukkalapalem, Ongole, Prakasam Dist, Affliation to Jawaharlal Nehru Technological university,kakinada in 2015-19 respectively.

Ms. **Garikapati Jaswitha** pursuing B Tech from computer science engineering in Qis college of Engineering and Technology(Autonomous & NAAC 'A' Grade), Pondure Road, vengamukkalapalem, Ongole, Prakasam Dist, Affliation to Jawaharlal Nehru Technological university,kakinada in 2015-19 respectively.

Ms. **Boina Aneesha** pursuing B Tech from computer science engineering in Qis college of Engineering and Technology(Autonomous & NAAC 'A' Grade), Pondure Road, vengamukkalapalem, Ongole, Prakasam Dist, Affliation to Jawaharlal Nehru Technological university,kakinada in 2015-19 respectively.

Ms. **Chillapalli Chandini** pursuing B Tech from computer science engineering in Qis college of Engineering and Technology(Autonomous & NAAC 'A' Grade), Pondure Road, vengamukkalapalem, Ongole, Prakasam Dist, Affliation to Jawaharlal Nehru Technological university,kakinada in 2015-19 respectively.

Mr. **Nannurisuresh** has received his M.Tech from QIS College of engineering and technology, Ongole, affiliated to JNT University Hyderabad in 2006. He is working as Assistant Professor in the department of Information Technology for the past 16 years and guided 9 P.G Students and 40 U.G students. His area of interest is Data Science.