Ensemble Based Utterance Level Multimodal opinion Mining Framework Ensemble Ensemble

V J Aiswaryadevi[1], S Kiruthika[2], G Priyanka[3], Sathyabama. S[4] ,N Nataraj[5]

[1]Assistant Professor,Department of Computer Science and Engineering,Sri Krishna College of Technology,Coimbatore. *Email: Aiswarya.devi@live.com.*

[2]Assistant Professor,Department of Computer Science and Engineering,Sri Krishna College of Technology,Coimbatore. *Email: Kiruthika.s@skct.edu.in.*

[3]Assistant Professor,Department of Computer Science and Engineering,Sri Krishna College of Technology,Coimbatore. *Email: Priyanka.g@skct.edu.in.*

[4]Assistant Professor, Sri Krishna College of Technology , PH-9500341062. E-mail: sathyabama.s @skct.edu.in

[5]Assistant Professor, Department of Information Technology,Bannari Amman Institute of Technology,Erode.*Email:Nataraj08@gmail.com*

Article Info
Volume 83
Page Number: 3736 - 3741
Publication Issue:
March - April 2020

Article History
***Article Received**: 24 July 2019*
***Revised**: 12 September 2019*
***Accepted**: 15 February 2020*
***Publication**: 23 March 2020*

Abstract:

Ensemble based algorithms are widely used for eradicating the issue of overfitting in classification algorithms and the framework proposed is reducing the issue of overfitting by bootstrapping the random samples and feature samples derived from the YouTube trending videos. The utterance level preprocessing on the video frames are extracted using the label encoder algorithm and the preprocessed data is fed into the ensemble-based regression model for extracting the sentiment features using random sampling and feature sampling. Random forest algorithm is one of the most efficient ensemble algorithms for reducing overfitting and space effective out of bag error elimination prediction model used to construct the proposed opinion mining framework. The YouTube trending videos are extracted with its linguistic information and MFCC is used to interpret the audio signals extracted from the same using APP algorithms. The weighted utterance level fusion of audio and linguistic information is attained with an accuracy of 88.29%. The fusion level of utterances with the image frame set and linguistic information is preprocessed using label encoder and OpenCV commands which attained the accuracy of 72.33% of sentiment polarity obtained. The fusion at the utterance level of linguistic, acoustic and visual modalities achieved the sentiment classification accuracy of 90.06%. The proposed opinion mining framework works with 91% accuracy on sentiment polarity identification from the emotion expressed in the videos fed as input.

Keywords: *Random sampling, Feature extraction, sentiment polarity, opinion mining, ensemble technique, Random forest, bootstrapping.*

I INTRODUCTION

The mean square error optimization is the main benefit of classification algorithms used for constructing opinion mining framework and it is positively achieved using ensemble based random forest algorithm and the accuracy is reported for proposed prediction model for discovering the emotional level in the facial features, gestures, acoustic signals and linguistic speech utterances delivered by the human generations and nowadays lot of classification models and opinion mining frameworks are developed and servicing the health care field with the effective solution for depression detection, Alzheimer disease detection and physically impaired people. Ensemble based algorithms are used for implementing the framework on the top of multiple trees with the elimination of overfitting and out of bag error precision.

The future of opinion mining framework is to embed the human emotions into Robots and the emotion recognition by the machines on the human beings. Section 2 represents the data preprocessing algorithms and their related prior work in applying those algorithms. Section 3 represents the random forest module replacing the deep learning algorithm with space efficiency and Section 4 represents the opinion mining framework developed out of ensemble algorithms. Results and future scope of the proposed framework will be discussed in the further sections of this paper.

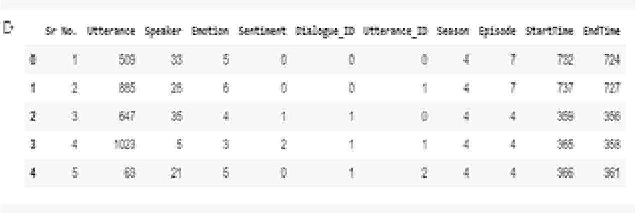
II Data Preprocessing Techniques

2.1 Related Work

IEMOCAP is pre-processed using KNN and fed into the ELM with SVM nonlinear prediction models and acquired with the prediction accuracy of 54% only in [1]. [Bitouk.et.al, 2010] constructed Hidden Markov Models (HMM) to differentiate the type of emotion has made use of utterance-level statistics, Such as Linear Prediction Coefficients (LPC) and Mel-Frequency Cepstral Coefficients (MFCC), have received less attention in emotion recognition Gaussian Mixture Models (GMMs) are commonly used to Image recognition. In our work we used two databases of emotional speech: an automatic emotional speech database from Linguistic Data Consortium (LDC) and Berlin database of German emotional speech. The prediction accuracy of speaker-independent emotion recognition. In this paper to say that the emotion recognition from speech signals using phoneme-class dependent HMM classifiers with short-term spectral features. [2] also calculated the accuracy rate in each specific phoneme class for 80%. The prediction of accuracy level of 55% and relative value was 0.7. [3] extracts the emotion level of speaker from the speech signals using phoneme-class dependent HMM classifiers with short-term spectral features. [Lee., 2004] also calculated the accuracy phenomena in each specific phoneme class for 80%. The prediction of accuracy level of 55% and relative value is 0.7%. In [4] author depicts that speech contain Emotion, Sub sentence, Segment and Decision model and the main database is constructed using the Berlin Emotional Speech Database (EMODB) (in German).The accuracy level of the performance gain using time based segment units with GMM decision model was attained as 3.5% and 4.2% (absolute) on USC and EMODB data respectively. [5] presents Speech emotion recognition using recurrent neural network, deep neural network, long short-term memory. Each emotional state using Gaussian mixture model (GMM) or hidden Markov model (HMM) is to be calculated efficiently by using a dynamic programming used to back-propagation through time (BPTT). The weighted accuracy of the proposed emotion recognition system was improved up to 12% compared to the DNN-ELM. The results obtained in indicate that using GMMs and three features level such as 1) standard MFCCs 2) MFCC-low using filters from 20 Hz to 300 Hz 3) pitch. A model was trained using Expectation Maximization (EM) algorithm. Performance is measured as absolute accuracy level in the range of 60% to 62%. Mao and Chen in 2009 proposed a new approach for emotion recognition based on a hybrid of hidden Markov models (HMMs) and artificial neural network (ANN). [7] have been using the three different types of database for Adopting Beihang University Database of Emotional Speech (BHUDES) and Berlin database of emotional speech, comparison between isolated HMMs and hybrid of HMMs/ANN.Feature Extraction of Mel prediction cepstrum coefficient (MPCC) and linear prediction cepstrum coefficient (LPCC) for two features to distinguish certain emotions. Emotional speech which was accurately recognized by at least 70 % of the listeners was collected into the experiment corpus by [7].[8] analyzed Performance, Experimentation, Design, Human Factors. The database used in the experiments was recorded from VICON. The author was concerned in obtaining an accuracy ranging from 70% to 92%. [Pérez-Rosas., 2013] automated the xtraction of sentiment from speech signals using MOUD data set and OpenEAR tool to compute a set of acoustic features. [9] has shown the accuracy level of 78% in emotion recognition from speech. In [10] the author in order to provide results on a public corpus decided for the Berlin Emotional Speech database (EMO-DB) and computed its accuracy for each classifier when the feature set is optimized by SVM-SFFS. Which led to be 93.6% accuracy on EMO-DB. IEMOCAP is preprocessed using Hidden Markov Models (HMM) and Dynamic Bayesian Networks (DBN)and acquired with the prediction accuracy of 65.15%. VICON methodology in [12] is used to predict the emotion level and the model was built using he Maximum Likelihood Bayes classifier (MLB), Kernel Regression (KR), and K-nearest Neighbors (KNN) with 89% of accuracy level. Walter proposed the hidden Markov automata (HMA) model in [13]to predict the emotion level with accuracy Level 72.2%. [Kim, et al.] in 2013 constructed an automatic emotion classification system with the prediction accuracy of 64% obtained through Multi-Dimensional Dynamic Time Warping (MD-DTW) K-NN model for IEMOCAP database. IEMOCAP was pre-processed using KNN and fed into the ELM with SVM nonlinear prediction models and acquired with the prediction accuracy of 82% only. GMM have been trained with the using this Expectation Maximization (EM) algorithm in [18] and the performance was measured as absolute accuracy level is 60%.

2.2 Data Pre-processing

Label encoder algorithm is used for linguistic content preprocessing and Mel Frequency Cepstral Coefficient (MFCC) is used for preprocessing the speech signals into time series data and the continuous linear regression is applied on the data for better accuracy and reduced mean square error of the sentiment content classified.



	Sr No.	Utterance	Speaker	Emotion	Sentiment	Dialogue_ID	Utterance_ID	Season	Episode	StartTime	EndTime
0	1	509	33	5	0	0	0	4	7	732	724
1	2	885	28	6	0	0	1	4	7	737	727
2	3	647	35	4	1	1	0	4	4	359	356
3	4	1023	5	3	2	1	1	4	4	365	358
4	5	63	21	5	0	1	2	4	4	366	361

Figure 1. Utterance level feature extraction from the YouTube video data set after preprocessing algorithm label encoder applied

Speech signal after preprocessing is taken into the test labels and feature list for driving the prediction accuracy.

[1 1 1 0 0 1 2 1 0 0 0 0 2 2 0 0 0 1 0 1 0 1 1 0 0 1 2 0 0 0 0 0 1 1 1 1 2

0 2 0 2 0 1 1 0 1 0 1 0 2 2 1 0 0 1 2 0 0 0 1 0 0 1 1 1 0 2 2 0 2 2 1 2 2

1 0 1 0 0 1 0 0 1 2 0 2 1 1 1 0 1 2 1 1 2 1 0 1 2 0 1 0 2 1 0 0 0 1 0 2 1

1 1 1 0 1 1 0 0 2 0 2 1 2 0 2 1 0 0 1 0 1 1 0 1 0 1 1 1 1 0 0 1 2 2 1 0 0

1 1 1 2 2 1 0 0 0 1 1 1 0 0 2 2 0 2 1 2 2 0 1 0 0 2 1 0 2 1 2 1 1 1 0 2 1

1 0 0 0 0 0 1 1 1 0 1 1 1 1 1 1 1 2 1 1 2 0 1 0 0 0 1 0 0 0 2 0 1 0 0 0 2

2 1 2 1 1 0 2 1 1 1 1 1 0 1 2 2 0 0 0 2 2 0 0 1 0 2 1 1 1 1 0 1 0 1 2 2 2

0 2 0 0 1 1 0 0 1 1 1 1 1 1 0 0 1 1 0]

III RANDOM FOREST CLASSIFIER

All the decision trees extracted in the form small subsequence tree are taken for bootstrapping and seeding with the seed rate of 100. Bootstrapped data set is used for generating the tree with minimum MSE specified in the labels of the nodes below.

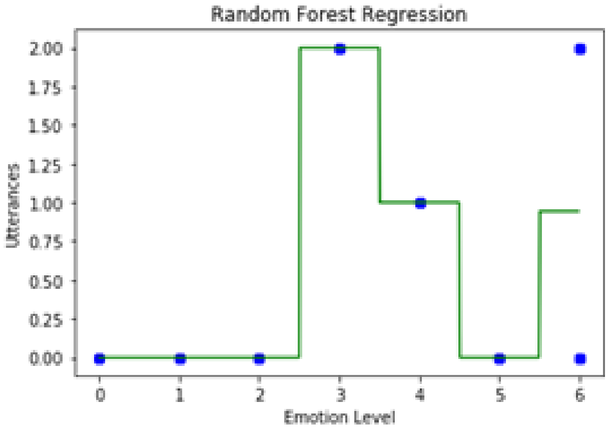


Figure 2. Random Forest Emotion classification in the Utterance level

The Markov process simply assumes that the "future is independent of the past given present. Angry, bored, neutral, happy, sad are the hidden states to be explored. Generative sequence model according to Naïve Bayes is to first decompose probability using Baye's law.

$$\text{argmax } P(y/x) = argmax\, P(X|Y)P(Y) \text{ ------ (1)}$$

In the next step, apply the HMM word emission probability for each POS tags identified iteratively.

$$P(X|Y) \approx \prod_1{}^{l} PE(Xi|Yi) \text{ ---------(2)}$$

Where l is the length of the sentence in corpus retrieved from the video sequence by Gaussian model. Ensemble based ELM model is used reduce the time taken by preprocessing on video signals. Smoothing is not necessary in case of HMM transition probability since there are many tags.

$$P(Yi|Yi\text{-}1) = PML(Yi|Yi\text{-}1) \text{ ------ (3)}$$

HMM emission probability is used for smoothening the unknown words as given below:

$$PE(Xi|Yi) = \lambda\, PML(Xi|Yi) + (1\text{-}\lambda)\ 1/N \text{---------(4)}$$

Feature scaling using MFCC is constructed using the time dependent framework derived from the naïve bayes rule.

IV OPINION MINING FRAMEWORK

The polarity of positive and negative feelings are classified using the opinion mining architecture.

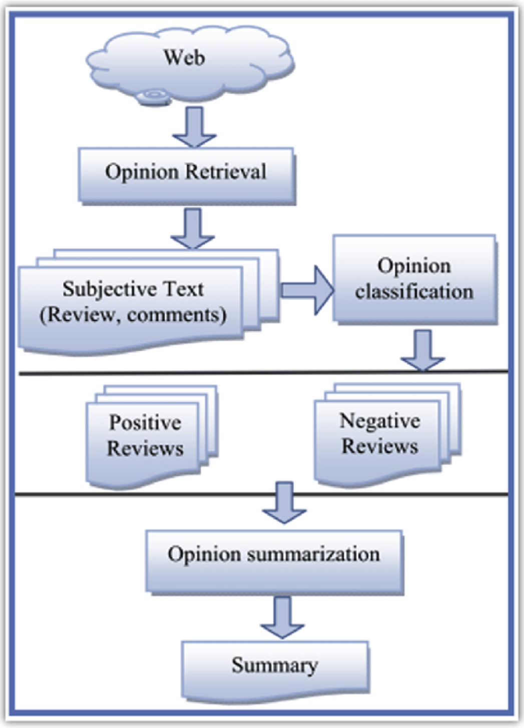


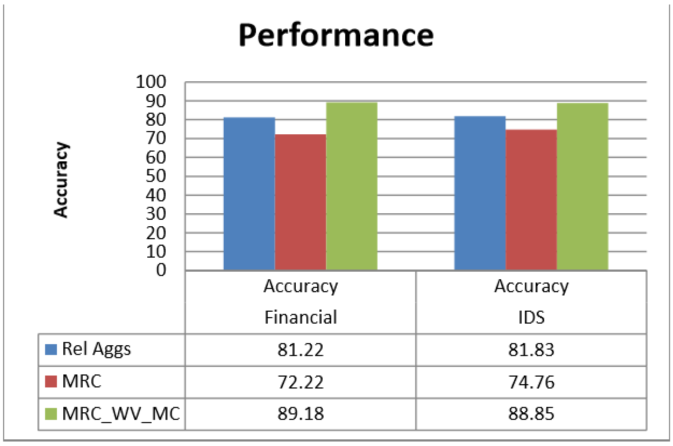
Figure 3. Opinion Mining Framework and Architecture

An unbiased opinion mining is useful in the health care field for prevention of health diseases in human Activities in Daily Life (ADL). The opinionated tokens are preprocessed as shown in the output of Fig 1 and the numerical, categorical data is used for random classifier prediction module. Positive opinions and negative opinions are used for ranking medical prescriptions for patient recovery in the prescribed duration. The growth of unstructured data sources, such as online reviews, posts and social media conversations, and structured data sources, such as stakeholders' letters, strategic documents and patents, did not generate a better understanding of the reality in which they operate, but rather made more complex the work of analysts. The following features allows the opinion mining framework to classify the polarity of emotions to a wide range of applications.

- opinion analysis: analyze opinions related to events or facts also under way, even when they are not related to specific topics;
- features extraction: extract aspects and significant information contained in the opinions, related to different contexts not always well defined, starting from multiple sources of reviews;
- domain specific features: contextualize the features through the use of tools for the semantic classification, the management of semantic networks, and the use of ad-hoc linguistic resources;
- reporting (opinion summarization): aggregate and represent the processed results so that they become useful information in decision making.

4.1 Results

88.29 % of accuracy is obtained by the fusion level of linguistic and speech features from the YouTube data set and 90.6% of accuracy is derived for the fusion of linguistic, acoustic and visual features of the data under consideration. The results derived from multiple modalities are represented below using Fig 4.



	Accuracy Financial	Accuracy IDS
🟦 Rel Aggs	81.22	81.83
🟥 MRC	72.22	74.76
🟩 MRC_WV_MC	89.18	88.85

Visual + Speech + Text

Figure 4. Performance Analysis of the Opinion Mining Framework

V Conclusion

The prediction models proposed in this paper is highly benefitting the multimodal emotion recognition framework through the opinion mining framework achieving high level of accuracy compared to any other machine learning algorithms. Large margin L-Softmax loss function is efficient deep neural networks rather than Euclidean distance loss. Fusion of data under multimodalities are gained with higher accuracy of sentiment derived. But neural networks and deep learning networks are incurring high level of mean square error to the maximum 0.18 but in random forest 4.67% error rate is identified and maximum out of bag error is located for highly accurate prediction model.

VI REFERENCES

[1] Han, K., Yu, D. and Tashev, I., 2014. Speech emotion recognition using deep neural network and extreme learning machine. In Fifteenth annual conference of the international speech communication association.

[2] Bitouk, D., Verma, R. and Nenkova, A., 2010. Class-level spectral features for emotion recognition. Speech communication, 52(7-8), pp.613-625.

[3] Lee, C.M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S. and Narayanan, S., 2004. Emotion recognition based on phoneme classes. In Eighth International Conference on Spoken Language Processing.

[4] Jeon, J.H., Xia, R. and Liu, Y., 2011, May. Sentence level emotion recognition based on decisions from subsentence segments. In 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 4940-4943). IEEE.

[5] Lee, J. and Tashev, I., 2015. High-level feature representation using recurrent neural network for speech emotion recognition. In Sixteenth Annual Conference of the International Speech Communication Association.

[6] Neiberg, D., Elenius, K. and Laskowski, K., 2006. Emotion recognition in spontaneous speech using GMMs. In Ninth international conference on spoken language processing.

[7] Mao, X., Chen, L. and Fu, L., 2009, March. Multi-level speech emotion recognition based on HMM and ANN. In 2009 WRI World congress on computer science and information engineering (Vol. 7, pp. 225-229). IEEE.

[8] Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Lee, S., Neumann, U. and Narayanan, S., 2004, October. Analysis of emotion recognition using facial expressions, speech and multimodal information. In Proceedings of the 6th international conference on Multimodal interfaces (pp. 205-211). ACM.

[9] Pérez-Rosas, V., Mihalcea, R. and Morency, L.P., 2013, August. Utterance-level multimodal sentiment analysis. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 973-982).

[10] Schuller, B. and Rigoll, G., 2006. Timing levels in segment-based speech emotion recognition. In Proc. INTERSPEECH 2006, Proc. Int. Conf. on Spoken Language Processing ICSLP, Pittsburgh, USA.

[11] Wöllmer, M., Metallinou, A., Eyben, F., Schuller, B. and Narayanan, S., 2010. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In Proc. INTERSPEECH 2010, Makuhari, Japan (pp. 2362-2365)

[12] Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Lee, S., Neumann, U. and Narayanan, S., 2004, October. Analysis of emotion recognition using facial expressions, speech and multimodal information. In Proceedings of the 6th international conference on Multimodal interfaces (pp. 205-211). ACM

[13] Walter, S., Scherer, S., Schels, M., Glodek, M., Hrabal, D., Schmidt, M., Böck, R., Limbrecht, K., Traue, H.C. and Schwenker, F., 2011, July. Multimodal emotion classification in naturalistic user behavior. In International Conference on Human-Computer Interaction (pp. 603-611). Springer, Berlin, Heidelberg

[14] Hazarika, D., Poria, S., Mihalcea, R., Cambria, E. and Zimmermann, R., 2018. ICON: interactive conversational memory network for multimodal emotion detection. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 2594-2604)

[15] Kim, Y. and Provost, E.M., 2013, May. Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 3677-3681). IEEE.

[16] Cho, J., Pappagari, R., Kulkarni, P., Villalba, J., Carmiel, Y. and Dehak, N., 2019. Deep neural networks for emotion recognition combining audio and transcripts. arXiv preprint arXiv:1911.00432

[17] Tan, Z.X., Goel, A., Nguyen, T.S. and Ong, D.C., 2019, May. A multimodal LSTM for predicting listener empathic responses over time. In 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019) (pp. 1-4). IEEE

[18] Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A. and Cambria, E., 2019, July. Dialoguernn: An attentive rnn for emotion detection in conversations. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, pp. 6818-6825)

[19] Huddar, M.G., Sannakki, S.S. and Rajpurohit, V.S., 2019. Multi-level context extraction and attention-based contextual inter-modal fusion for multimodal sentiment analysis and emotion classification. International Journal of Multimedia Information Retrieval, pp.1-10

[20] Basha, Jameer & Purusothaman, T.. (2011). Efficient Multimodal Biometric Authentication Using Fast Fingerprint Verification and Enhanced Iris Features. Journal of Computer Science. 7. 698-706. 10.3844/jcssp.2011.698.706.

[21] Perumal, Tamije Selvy & Purusothaman, T. (2011). Performance Analysis of Clustering Algorithms in Brain Tumor Detection of MR Images. Eur J Sci Res. 62.

[22] Uhrmann, L.S., Nordli, H., Fekete, O.R. and Bonsaksen, T., 2017. Perceptions of a Norwegian clubhouse among its members: A psychometric evaluation of a user satisfaction tool. International Journal of Psychosocial Rehabilitation, 21(2).

[23] Das, B. and KJ, M., 2017. Disability In Schizophrenia and Bipolar Affective Disorder. *International Journal of Psychosocial Rehabilitation*, *21*(2).