An Ensemble Approach to Utterance Level Multimodal Sentiment Analysis

Mahesh G. Huddar
Dept. of Computer Science and Engineering
Hirasugar Institute of Technology, Nidasoshi, Belagavi
Karnataka, India
mailtomgh1@gmail.com

Dr. Sanjeev S. Sannakki
Dept. of Computer Science and Engineering
Gogte Institute of Technology, Belagavi
Karnataka, India
sannakkisanjeev@gmail.com

Dr. Vijay S. Rajpurohit
Dept. of Computer Science and Engineering
Gogte Institute of Technology, Belagavi
Karnataka, India
vijaysr2k@yahoo.com

***Abstract* - The primary objective of sentiment analysis system is to automatically discover and analyze people's attitude, opinion, or position towards a product, a topic, a person or an entity. A huge amount of multimedia content is being posted on social websites such as YouTube, Flicker, and Twitter on every day. To cope up with such multimedia data, there is a need for state-of-the-art multimodal sentiment analysis framework that can extract information from multimodal data. The purpose of this research work is to improve the accuracy of sentiment prediction by analyzing the textual features along with facial expressions. We examine what people say and their facial expressions when they are saying it. Bag-of-words representation is used to create textual features. Facial expressions and audio features were extracted using open source tools such as OpenFace and OpenSmile respectively. Unimodal, bimodal, trimodal and ensemble approaches were used for classification. Our results demonstrate proposed ensemble approach outperforms other base models.**

Keywords: Multimodal Sentiment Analysis, Computer Vision, Machine Learning, Bag-of-Words, Ensemble approach

I. Introduction

Automatic sentiment analysis is a process of uncovering one's opinion about a topic or entity [1]. Extracting people's opinion towards a certain topic, entity or person has many applications. For example, Political parties are interested in knowing the opinion of voters to gauge the voting intensities [2]. Companies are interested in understanding the customer's opinion regarding their product or brand [3]. With the emergence of social and World Wide Web (WWW), individuals started expressing their opinion, position or attitude towards a topic or entity through these media. This leads to the humongous and rich amount of data set for sentiment analysis and the development of techniques for automatic sentiment analysis [4]. Initially, sentiment analysis is restricted to text based analysis and only recently sentiment analysis and emotion detection from other modalities such as audio, video and image, begun considered. The user-generated data available on social media, containing people's position or opinion, is noisy and unstructured in nature. Also, posts with ambiguous, irony or negation phrases with implicit meaning add new challenges to automatic sentiment analysis system [5]. Text based sentiment analysis is a well-researched topic which has a lot of applications in various domains such as stock market performance prediction [6], election outcome prediction [7], movie box office performance prediction [8].

In the remainder of this paper, we discuss related work in section 2. Section 3 describes the dataset and feature extraction from the dataset. Proposed multimodal sentiment analysis methodology is discussed in Section 4. Proposed ensemble approach is discussed in section 5. Finally, results obtained from base classifiers, traditional ensemble approaches, and proposed ensemble approach are discussed in section 5 and conclude in section 6.

II. Litarature Review

Initially, sentiment analysis started as an alternative to topic detection and tracking. Back in early 2000, numerous works were carried to address the problem of sentiment analysis from textual data. The influential review [4] in 2008 increases the interest among researchers in this field, and subsequently, new methods, approaches, and applications have been developed in recent research [9]. Significant studies have been carried out to detect positive, negative, or neutral sentiment associated with words [10], multi-words [11], phrases [12], sentences [13], and documents [14]. Today we can witness a shift from textual social web to multimodal social web. For example, users post their opinion, position or attitude in the form of images on Instagram, flicker, Twitter along with a textual tag and post spoken and video reviews on YouTube. Currently, research in multimodal sentiment analysis focuses on analyzing sentiment from images and the tag associated with images that are Visual sentiment analysis [15] and Sentiment analysis from audio-visual content [16] [17] [18]. Recently many researchers have attempted to discover the opinion expressed in the social web from multimodal content, including textual, audio and visual information. Usually, multimodal fusion is performed either at feature level (early fusion) or at decision level (late fusion). In feature level fusion multimodal features such as textual features and audio-visual features are concatenated at the input level and the final classification is performed on combined features [19]. In decision level fusion, each of the multimodal features is individually classified and the results of each classifier are fused to get the final result [20] [21]. Numerous studies demonstrated that speech analysis can be used for sentiment analysis [22] [23]. There are several surveys on textual [24] and multimodal [25] sentiment analysis. Researchers have used different approaches to address the problem of multimodal sentiment analysis.

III. Data Set and Feature Extraction

CMU-MOSI "Multimodal Opinion Sentiment Intensity" dataset is used in this work [18]. This is a sentiment annotated English movie reviews dataset which is collected from YouTube. There are 93 distinct speakers in CMU-MOSI with 2199 opinion utterances. Open source software OpenFace [26] toolkit is used to extract facial expressions such as smile intensity, the distance between eyes, nose position, head poses, head shake, frown and gaze direction. OpenSMILE [27] open source software was used to extract acoustic low-level descriptors (LLD). It includes voice quality, prosodic and Mel Frequency Cepstral Coefficients (MFCC). A total of 991 features were extracted using the OpenSmile software. Using subset selection technique 52 features were selected such as intensity, loudness, MFCC, LSP frequency, PCM, PCM intensity, PCM loudness, and voice probability. Then the mean and standard deviation were calculated which makes a vector of 104 audio features for our experimental analysis. For textual analysis, it was transcribed using automatic speech recognition software. TF-IDF textual features were extracted from the textual content.

IV. Implementation

MOSI dataset consists of 2199 opinion utterances. Among these 1751 utterances were selected for training the classification models. 458 were used for testing the classification models. Following are the steps followed in the implementation of multimodal sentiment analysis.

Step 1. In the first step videos reviews were segmented into opinion level utterances. The average length of utterance in 4.2 seconds and on average consists of 12 words.

Step 2. Feature Extraction for MOSI dataset.

- a. Textual feature extraction
- b. Feature extraction from audio using the OpenSmile toolkit.
- c. Feature extraction from video using the OpenFace toolkit.

Step 3. Features were normalized using min-max transformations.

Step 4. Subset selection method is applied to select features from audio and video.

Step 5. Multimodal feature vector is prepared by concatenating unimodal feature vectors (feature level fusion).

Step 6. Dataset is divided into training dataset (consisting of 1751 utterances) and testing dataset (consisting of 458 utterances).

Step 7. Sentiment classification based on base classifiers for different combinations of modalities such as only text, only audio, only video, only text and audio, only text and video, only audio and video and all three modalities,

- a. Support Vector Machines (SVM) – It is also called probabilistic classifier [28]. Training time of SVM is slow but it is more accurate compared to other models. The parameter values of the SVM classifiers are tuned as: C = 1, gamma = 1 and kernel = linear.
- b. Linear Regression – This is a regression model, which can also be used for classification. LR is generally used to relate a single categorical dependent variable to one or more independent variables [29]. The parameter values of the LR classifiers are tuned as: C = 0.01, maxiter = 80
- c. Decision tree – Decision tree classifier provides a hierarchical decomposition of the training dataset space in which a condition on the attribute value is used to divide the dataset [30]. The parameter values of the RF classifiers are tuned as: maxdepth = 30.
- d. K – Nearest Neighbor (KNN) – KNN is very popular and easy to implement machine learning algorithms for text classification [31]. The main objective of K-NN algorithm is to classify data into one of the predefined. The parameter values of the KNN classifiers are tuned as: nearest neighbor=20.
- e. Random forest – Random forest is a form of ensemble approach [32]. Random forest is a combination of decision tree classifier. The parameter values of the RF classifiers are tuned as: nestimators = 125, maxdepth = 20.

Step 8. Next proposed ensemble approach in designed and implemented.

Step 9. Performance of proposed approach is compared against traditional ensemble models such as voting, averaging and optimal weighted averaging.

The steps in multimodal sentiment analysis are shown in in figure 1.

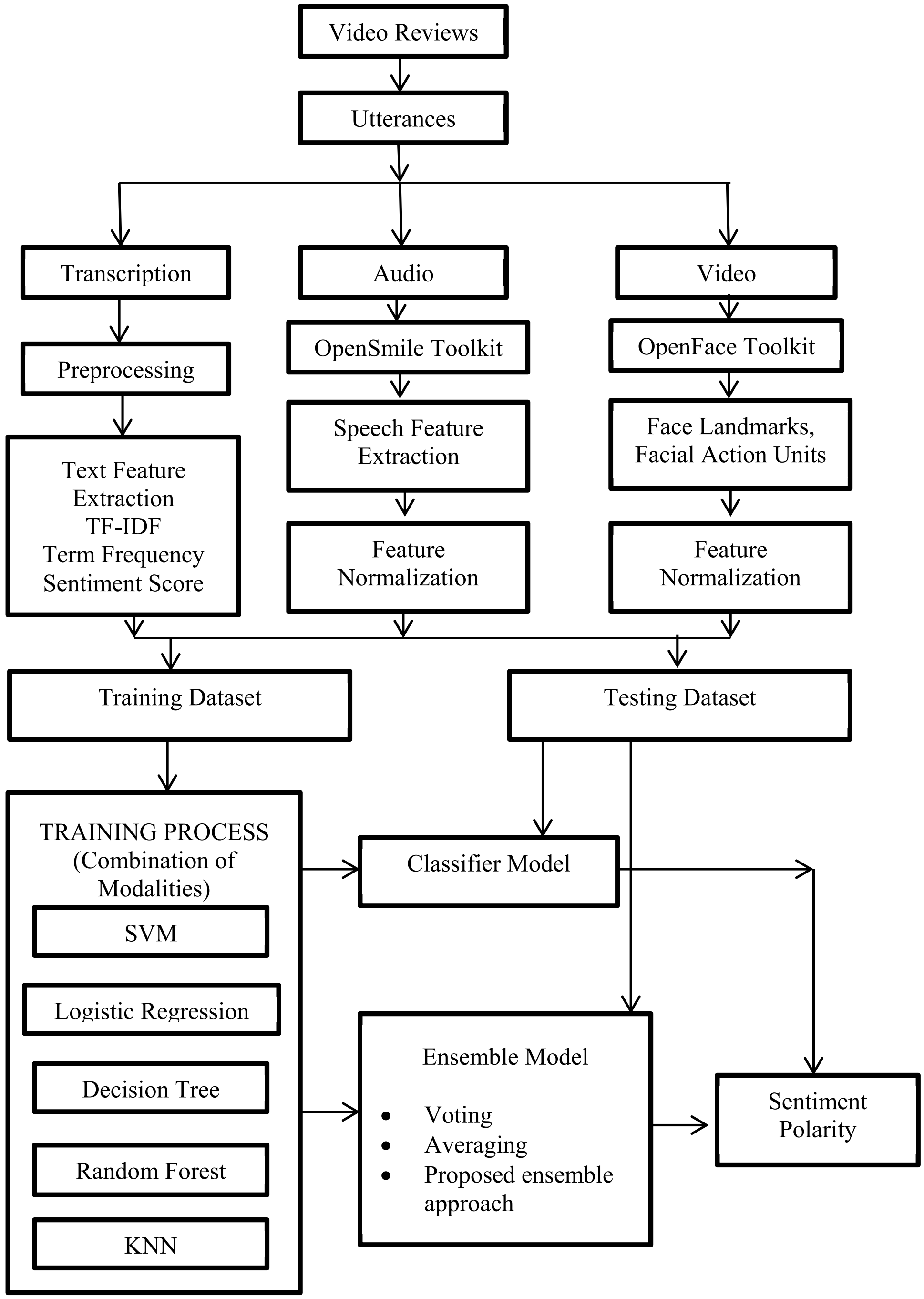

Figure 1. Methodology of multimodal sentiment analysis

V. PROPOSED ENSEMBLE CLASSFIER

Proposed algorithm calculates sentiment score and sentiment label of the test review. The proposed model was trained using the training dataset. The test dataset is used to test the model. Each of the base classifier determines the sentiment (Positive/Negative) of each review in the test dataset. The next step is to calculate the probability of each review being positive and negative. After assigning classification probability, each classifier in the ensemble technique is assigned with a weight based on the accuracy of each classifier. Then the positive and negative score of the testing review is calculated using sentiment probability and the weight assigned to the classifiers. If positive score is greater than negative score then testing review assigned positive sentiment otherwise negative sentiment.

Algorithm – Proposed ensemble algorithm to calculate the sentiment score and sentiment of a test review

$\boldsymbol{Function\ to\ Calculate\ Sentiment\ score\ (Test\ review)}$

$Input:\ Test\ review$

$Output: Sentiment\ Score$

$foreach\ Reviewi\ in\ Testreview$

do

$PositiveCounti = 0$

$NegativeCounti = 0$

$foreach\ classifier\ Ci\ in\ classifier\ ensemble$

do

$if\ Ci\ predict\ Positive$

$then$

$PositiveCounti += 1$

$else$

$NegativeCounti += 1$

end

end

$$Probability(Positivei) = \frac{PositiveCounti}{PositiveCounti + NegativeCounti}$$

$$Probability(Negativei) = \frac{NegativeCounti}{PositiveCounti + NegativeCounti}$$

end

$foreach\ classifier\ Ci\ in\ classifier\ ensemble$

do

$$WeightCi = \frac{AccuraccyCi}{\sum_{j=1}^{n} AccuraccyCj}$$

end

$Where\ AccuraccyCi\ represents\ Accuracy\ of\ Cith\ classifier$

$Where\ AccuraccyCj\ represents\ Accuracy\ of\ Cjth\ classifier$

$foreach\ reviewi\ in\ Test\ review$

do

$PositiveScorei = 0$

$NegativeScorei = 0$

$foreach\ classifier\ Ci\ in\ classifier\ ensemble$

do

$if\ Ci\ predicts\ Positive$

$then$

$PositiveScorei += WeightCi * Probability(Positivei)$

$else$

$NeativeScorei += WeightCi * Probability(Negativei)$

end

end

end

$\boldsymbol{Function\ to\ calculate\ sentiment\ predictor}$

$Input: PositiveScorei, NegativeScorei$

$Output: Sentiment\ Label$

$If\ Positivescorei > Negativescorei$

$then$

$Sentiment = \text{“Positive”}$

$else\ if\ Negative\ scorei > Positive\ scoreithen$

$Sentiment = \text{“Negative”}$

$else$

Find the most similar review in the testing dataset, using cosine similarity of testing $reviewi$ *with all other reviews in test dataset using formula 1. Let's say most similar review is* $reviewj$.

Calculate $PositiveScorej$ *and* $NegativeScorej$ *of review using function sentiment score*

$if\ Positive\ scorej >= Negative\ scorej$

$then$

$Sentiment = \text{“Positive”}$

$else$

$Sentiment = \text{“Negative”}$

end

$return\ Sentiment$

end

Cosine Distance Calculation: “Cosine similarity” measures the similarity of a pair of reviews [33]. Cosine similarity can be computed by using the following formula:

$$cosine\ Similarity(review1, review2) = \frac{review1 \,.\, review2}{||review1|| \cdot ||review2||} \quad ----1$$

where review1 and review2 represent feature vectors.

VI. EVALUATION AND RESULTS

The proposed system was tested on the CMU-MOSI dataset. Initially, base classifiers were tested for different modalities. Results were shown in table 1. Results show that text-audio, test-video, and text-audio-video perform better compared to other modalities.

Table 1. Comparison of results obtained from base classifiers.

Classifier → Modality ↓	SVM	LR	KNN	RF	DT
Text	0.727	0.747	0.706	0.654	0.626
Audio	0.62	0.631	0.62	0.597	0.515
Video	0.597	0.592	0.64	0.576	0.563
Text-Audio	0.759	0.761	0.708	0.554	0.572
Text-Video	0.747	0.772	0.72	0.677	0.633
Audio-Video	0.62	0.642	0.608	0.588	0.588
Text-Audio-Video	**0.759**	**0.774**	**0.729**	**0.681**	**0.651**

Figure 2 show the comparison of results obtained from base classifiers.

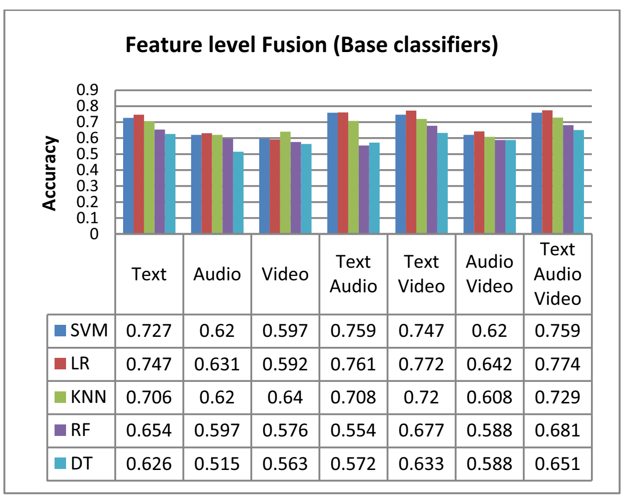

Figure 2. Comparison of base classifier results

Then results of base classifiers are used to obtain results from traditional ensemble approach and proposed ensemble approach. The results are shown in table 2. The result demonstrates that traditional ensemble approaches outperform base classifiers. Proposed ensemble approach performs better than base classifiers and traditional ensemble approaches.

Table 2. Comparison of results obtained from an ensemble approach.

Ensemble Approach → Modality ↓	Voting	Averaging	Proposed Ensemble approach
Text	0.743	0.722	0.753
Audio	0.638	0.554	0.658
Video	0.613	0.569	0.641
Text-Audio	0.761	0.688	**0.786**
Text-Video	0.779	0.743	**0.781**
Audio-Video	0.663	0.61	0.682
Text-Audio-Video	0.772	0.749	**0.797**

Figure 3 shows the comparison of results obtained from traditional and proposed ensemble approaches.

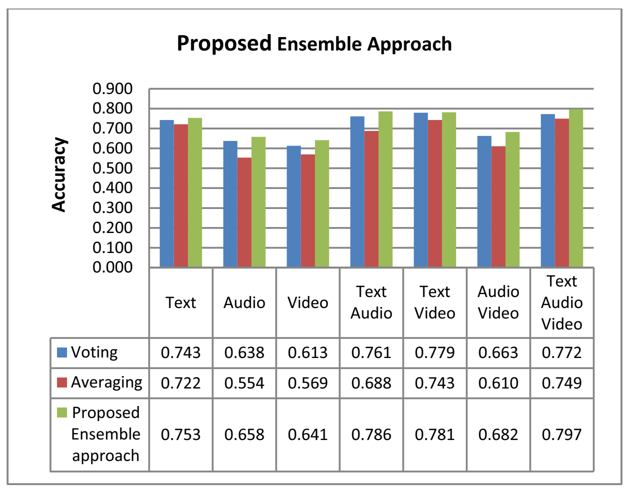

Figure 3. Comparison of base ensemble approach results

VII. CONCLUSION

The basic multimodal sentiment analysis approach is to compare the different base classifiers and select the best among them. The ensemble based classification approaches were being used widely in many areas to solve machine learning problems. In this paper, an ensemble based classification model has been proposed. The performance of the proposed ensemble model is compared against base classifiers and traditional ensemble approaches such as voting and averaging. The proposed ensemble classification model is formed by different base classifiers like Support vector machines, Random Forest classifier, Decision trees, K Nearest Neighbor, and Logistic Regression. The results show that the proposed ensemble approach outperforms standalone base classifiers and the traditional ensemble classifier. As future work, need to use neural networks based approaches which are most likely to perform better than the base classifier and traditional ensemble approaches.

REFERENCES

[1] L. Z. Bing Liu, "A survey of opinion mining and sentiment analysis," in *Mining Text Data*, Springer US, 2013, pp. 415-463.

[2] M. Prem, G. Wojciech and R. D. Lawrence, "Sentiment analysis of blogs by combining lexical knowledge with text classification," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Paris, France, 2009.

[3] B. J. Jansen, M. Zhang, K. Sobel and A. Chowdury, "Twitter power: Tweets as electronic word of mouth," *Journal of the Association for Information Science and Technology,* vol. 60, no. 11, p. 2169–2188, 2009.

[4] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval,* vol. 2, no. 1-2, pp. 1-135, 2008.

[5] E. Cambria, B. Schuller and Y. Xia, "New Avenues in Opinion Mining and Sentiment Analysis," *IEEE Intelligent Systems,* vol. 28, no. 2, pp. 15 - 21, 2013.

[6] J. Bollen, H. Mao and X.-J. Zeng, "Twitter mood predicts the stock market," *Journal of Computer Science,* vol. 2, no. 1, pp. 1-8, 2011.

[7] A. Tumasjan, T. O. Sprenger, P. G. Sandner and I. M. Welpe, "Predicting elections with Twitter: what 140 characters reveal about political sentiment," in *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, George Washington University, 2010.

[8] S. Asur and B. A. Huberman, "Predicting the Future With Social Media," *Web Intelligence and. Intelligent Agent Technology (WI-IAT),* vol. 1, no. 6, pp. 492-299, 2010.

[9] F. N. Ribeiro, M. Araújo, P. Gonçalves, M. A. Gonçalves and F. Benevenuto, "SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods," *EPJ Data Science,* vol. 5, no. 23, pp. 1-29, 2016.

[10] P. D. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics*, Philadelphia,, 2002.

[11] E. Cambria, D. Olsher and D. Rajagopal, "SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis," in *Proceedings of AAAI'14 Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, Québec City, Québec, Canada, 2014.

[12] T. Wilson, J. Wiebe and P. Hoffmann, "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis," in *In: HLT/EMNLP*, Vancouver, BC, Canada, 2005.

[13] E. Riloff and J. Wiebe, "Learning extraction patterns for subjective expressions," in *EMNLP '03 Proceedings of the 2003 conference on Empirical methods in natural language processing*, Stroudsburg, PA, USA, 2003.

[14] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," in *Proceedings of the ACL*, Barcelona, Spain, 2004.

[15] D. Borth, R. Ji, T. Chen, T. Breuel and S.-F. Chang, "Large-scale Visual Sentiment Ontology and detectors using adjective noun pairs," in *ACM International Conference on Multimedia*, Barcelona, Spain, 2013.

[16] V. P. Rosas, R. Mihalcea and L.-P. Morency, "Multimodal sentiment analysis of Spanish online videos," *IEEE Intelligent Systems,* vol. 28, no. 3, pp. 38-45, 2013.

[17] M. Wöllmer, F. Weninger, T. Knaup and B. Schuller, "YouTube movie reviews: ssentiment analysis in an audio-visual context," *IEEE Intelligent Systems,* vol. 28, no. 3, pp. 46-52, 2013.

[18] A. Zadeh, R. Zellers and E. Pincus, "Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages," *IEEE Intelligent Systems,* vol. 31, no. 6, pp. 82 - 88, 2016.

[19] S. Poria, E. C. Iti Chaturvedi and A. Hussain, "Convolutional MKL Based Multimodal," in *IEEE 16th International Conference on Data Mining*, 2016.

[20] H. Wang, A. Meghawat, L.-P. Morency and E. P. Xing, "Select-Additive Learning: Improving Cross-individual Generalization in Multimodal Sentiment Analysis," *arXiv preprint,* 2016 .

[21] A. Zadeh, R. Zellers, E. Pincus and L.-P. Morency, "MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos," *IEEE Intelligent Systems,* vol. 31, no. 6, pp. 82-88, 2016.

[22] J. C. Pereira, J. Luque and X. Anguera, "Sentiment retrieval on web reviews using spontaneous natural speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[23] L. Kaushik, A. Sangwan and J. H. L. Hansen, "Automatic sentiment extraction from YouTube videos," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2013.

[24] A. M. Abirami and V. Gayathri, "A survey on sentiment analysis methods and approach," in *2016 Eighth International Conference on Advanced Computing (ICoAC)*, Chennai, India, 2017.

[25] M. Soleymani, D. Garcia, B. Jou and M. Pantic, "A Survey of Multimodal Sentiment Analysis," *Image and Vision Computing,* vol. 65, pp. 3-14, 2017.

[26] T. Baltrusaitis, A. Zadeh, Y. C. Lim and L. P. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," in *IEEE International Conference on Automatic Face & Gesture Recognition*, FG, 2018.

[27] F. Eyben, M. Wöllmer and B. Schuller, "openSMILE: The Munich versatile and fast open-source audio feature extractor," in *ACM International Conference on Multimedia (MM)*, 2010.

[28] A. S. Nurulhuda Zainuddin, "Sentiment analysis using Support Vector Machine," in *International Conference on Computer, Communications, and Control Technology (I4CT)*, Langkawi, Malaysia, 2014.

[29] W. P. Ramadhan, A. Novianty and C. Setianingsih, "Sentiment analysis using multinomial logistic regression," in *International Conference on Control, Electronics, Renewable Energy and Communications*, Yogyakarta, Indonesia, 2017.

[30] Y. Zhong, "The analysis of cases based on decision tree," in *IEEE International Conference on Software Engineering and Service Science* , Beijing, China, 2016.

[31] B. Trstenjak, S. Mikac and D. Donko, "KNN with TF-IDF Based Framework for Text Categorization," in *24th DAAAM International Symposium on Intelligent Manufacturing and Automation*, 2013.

[32] Jianqiang Z, "Combing semantic and prior polarity features for boosting twitter sentiment analysis using ensemble learnin," in *IEEE International Conference on DataScience in Cyberspace (DSC)*, 2016.

[33] N. Cheng, Z. Yu and K. Wang, "String similarity computing based on position and cosine," in *7th IEEE International Conference on Electronics Information and Emergency Communication (ICEIEC)*, Macau, China, 2017.