A User-oriented Ensemble Method for Multi-Modal Emotion Recognition

Nishan Iddamalgoda[1], Pavani Thrimavithana[2], Hiruni Fernando[3], Thilini Ratnayake[4], Y.H.P.P Priyadarshana[5], Rekha Aththidiye[6] and Dharshana Kasthurirathna[7]

[1,2,3,4,5,7]Faculty of Computing, Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
[6]Faculty of Graduate Studies and Research, University of Colombo
Colombo, Sri Lanka

[1]nishan.iddamalgoda@gmail.com, [2]yasarathrima@gmail.com,
[3]hira94.fer@gmail.com, [4]thilinirathnayake94@gmail.com, [5]prasan.y@sliit.lk,
[6]rekha.aththidiye@gmail.com, [7]dharshana.k@sliit.lk

Abstract. Emotions play a vital role in mental and physical activities of human lives. One of the biggest challenges in Human-Computer Interaction is emotion recognition. With the resurgence in the fields of Artificial Intelligence and Machine learning, a considerable number of studies have been carried out in order to address the challenge of emotion recognition. The individual heterogeneity of expressing emotions is a key problem that needs to be addressed in accurately detecting the emotional state of an individual. The purpose of this work is to propose a novel ensemble method to predict the emotions using a multimodal approach. The presented multimodal approach with the modalities of facial expressions, voice variations and, speech and social media content, are used to identify seven emotional states: anger, fear, disgust, happiness, sadness, surprise and neutral emotion. In this study, for the facial expression-based emotion recognition and voice variation-based emotion recognition, Deep Neural Network models have been used, and for emotion recognition using speech and social media content, Multinomial Naïve Bayesian algorithm is used. The mentioned three modalities were integrated using a novel ensemble method that captures the heterogeneity of individuals in how they express their emotions. The proposed ensemble method was evaluated with respect to real states of human emotions of a sample user group and the experimental results suggest that the suggested ensemble method may be more accurate in recognizing emotions. Accurate recognition of emotions may have myriad applications in domains such as healthcare, advertising and human resource management.

Keywords**:** emotion recognition, ensemble methods, deep learning, machine learning

1 Introduction

Emotions are psychological states that are generated subconsciously. Generally, they are autonomous body responses to certain external or internal events [1]. The word "emotion" has been originated from the Latin "ex" which means out and "motio" that means moving, so "emotion" indicates a movement, in particular, a body movement [2]. Scrutinizing of emotions has been a very active research area. In fact, the initial research on emotions can be found in the book "The Expression of the Emotions in Man and Animals" by Charles Darwin [3]. In 1971 Ekman and Friesen identified six basic emotions: happiness, surprise, fear, disgust, anger and sadness in the study on the universality of facial expressions [4].

Although everyone experiences these emotions; it is rather difficult to identify the emotions that are expressed or how they can be measured. This is particularly relevant when identifying and treating for emotional disorders. Emotional disorders are conditions that mainly involve pathological changes in emotions, thinking or behaviour (or a combination of all) [5]. Further, the recognized and quantified emotional state can be considered as a useful input in applications spanning multiple domains such as targetted advertising and mental healthcare. Multitude of data points can be considered for detecting the emotional state of an individual, such as the facial expressions, voice variations and speech content and texual expressions such as social media posts. In this work, we attempt to combine the predictions obtained using these multiple modalities, in order to give an aggregated value of the overall emotional state.

It is well established in psychology that each individual may have a unique way of expressing their emotions through these modalities [5], [6]. For instance, some may be more expressive through facial expressions or voice variations, while others may be more apt with using text to express their emotional state. Hence, a holistic ensemble method that tries to fusion the predictions given using multiple modalaities should take into account this heterogeneity of individuals in expressing one's emotional sttate.

In this work, we use the following modalities to identify the emotional state of an individual,

Recognizing the emotional state using facial expressions.

Recognizing the emotional state using voice variations such as the tone and pitch of the voice.

Recognizing the emotional state using social media content and speech content.

In the proposed approach, we employ an ensemble method to combine the predictions obtained using each modality and to provide an aggregated result. We introduce a novel, weight driven ensemble method that entails a user-specific learning process that attempts to apply a user-specific weight to each modality, depending on how each user expresses their emotional state through each modality. In the context of this work, a user is an individual whose emotional state that needs to be accurately recognized.

The next section discusses the relevant background of this work. Next, we present the methodology that we used to recognize the emotional state using multiple modalities and the evaluation method used. Then we present the results obtained in implementing the

proposed method and results of the experiments conducted for the evaluation of the proposed method. Finally, we present our conclusions with a brief description of the limitations of this work and a discussion on the potential future work.

2 Background

Emotions in human are expressed by different modes like facial expressions, body language and speech. These modalities, independently or in various combinations can be used for emotion recognition. Although a vast number of studies have been carried out independently for the techniques mentioned above, the multimodal approach to identify the emotions of a user is relatively less studied.

In the recent past, an extensive number of studies have been carried out for emotion recognition using facial features, speech features, social media and text content of users independently. S. Alizadeh and A. Fazel [11] have presented an approach based on a Convolution Neural Network (CNN) for Facial Expression Recognition (FER), where six emotions are recognized. K. Shan et al. [12] suggested an approach to recognize the facial expressions based on deep CNN. The proposed system consists of four-layer CNN architecture. In addition, a comparison between CNN and K Nearest Neighbour (KNN) algorithm is mentioned. The proposed system achieved a performance accuracy of 76.77% and 80.30% for Japanese Female Facial Expression (JAFFE) and Extended Cohn Kade (CK+) dataset respectively. Lopes et al. [13] proposed an FER system that uses a combination of CNN and specific image processing steps. The proposed method achieved an accuracy of 96.76% on the CK+ dataset.

A. M. Badshah et al. [14] presented an approach for Speech Emotion Recognition (SER) using deep CNN and spectrogram images that were generated from the Berlin Speech Emotion dataset. The proposed model achieved an accuracy of 84.3%, and aggregation method is used to combine the individual predictions into an overall prediction result for the entire audio file. The architecture proposed by E. Fran I. et al. [15] has been developed with a set of Romanian language recordings and an experimental software environment. The proposed architecture has been adapted by image processing and CNN techniques. The proposed model achieves the mean accuracy of 71.33% for six basic human emotions. Another SER system [16] has been implemented by utilizing Deep Neural Network (DNN) to recognize human speech emotion. The authors have chosen the Mel-frequency Cepstral Coefficients (MFCC) and then extracted speech features were fed into DNN for the training purpose of the network. Based on the accuracy rate MFCC, a number of neurons and layers are adjusted for optimization. The proposed system achieved a total recognition rate of 96.3% for three emotions and 97.1% for four emotions with the optimum configuration for SER.

S. Chaffar and D. Inkpen [17] have presented a supervised learning approach for the automatic emotion recognition from the text for the six basic emotions identified by P. Ekman [4]. For the presented approach four dataset have been use; and a comparison between the J48 and Naïve Bayes classification method have been conducted to identify the best classification algorithm. U. Jain and A. Sandhu [18] have presented a supervised learning approach comparing with unsupervised learning approaches. They have used lexical resources as features in machine learning algorithms to obtain good results. In this paper, they summarize the most kind of system use features based on shallow analysis of the text as n-grams, punctuations, emoticons or part of speech, which can also be considered as a non-multimodal approach.

There are some key studies that attempt to apply ensemble methods to solve the problem of emotion prediction. I. Perikos and I. Hatzilygeroudis have presented a classifier system based on an ensemble mechanism for textual sentiment analysis [19]. Here they have focused only on emotional textual contents such as articles, news headlines and mainly social media contents. Even though the ensemble technique has performed better than other mechanisms, still this is limited for textual emotional contents. Morrison et al. have introduced set of ensemble methods which can be used to evaluate speech emotion prediction [20]. They have evaluated natural and acted emotional acoustic contents and the experimental results revealed that the natural acoustic contents has always been performed well on the basis of ensemble speech classification methods such as StackingC and vote. The authors have suggested multi-modality as well but they have kept it as future enhancements.

In the literature on automated emotion recognition, there exists some work that attempt to use multiple modalities to identify and recognize the emotional state of a person. C. Zucco et al. [8], [9], [10] proposed several fusion approaches by the application of sentimental analysis and affective computing methodologies for emotion prediction and monitoring in order to integrate information extracted from multiple modes of the system. Kumar et al. [8] proposed a feature level fusion technique and decision level techniques in the multimodal data analysis system. I.T. Meftah et al. [9] proposed a multimodal approach to recognize the basic emotions and also combining emotions like simulated and masked emotions by combining modalities like facial expressions, speech expressions, gestures and body gestures. Cid et al. [10] proposed a multimodal emotion recognition system based on the visual and auditory information to detect five emotional states of humans.

F. Lingenfelser et al. [21] have researched on various multi-modal fusion techniques for audio visual emotion recognition. Their experiments have been based on facial and vocal modalities. But the textual context has not been taken into consideration. Padgett et al. [22] has presented an emotion recognition system based on six basic emotions. They have succeeded in introducing 12 ensemble independent network models for facial analysis. They have used a neural network model which consists of a multi-layer ensemble models. But still this approach also cannot be considered as a multi-modularity mechanism.

In summary, numerous studies have been conducted in the past decades about the systems that can monitor and assist emotions. However, in these studies, the number of modalities considered are fairly limited. In

particular, the multi-modal approach for emotion recognition has not been accounted well enough. Although,several ensemble methods have been used to aggregate the predictions obtained using multiple modalities, the individual heterogeneity in expressing emotions is not taken into account, which is the main limitation that we try to address in this work.

3 Methodology

The following subsections elaborate on the methods used for emotion recognition, using each input modality. Each input modality is input to a different model to predict the emotional state of the user and the resulting emotional predictions are aggregated using the proposed user-oriented ensemble method.

3.1 Emotion Recognition using Facial Expressions

Facial expressions provide a key avenue to recognize human emotions. It is a form of non-verbal communication, and they are the primary means of conveying social messages between humans. The general approach to facial emotion analysis consists of three steps: face detection, feature extraction and feature classification [23].

In this work, we employed a CNN model combined with image pre-processing techniques to identify/predict the facial emotions. The reason for selecting the CNN model was its effectiveness in extracting non-trivial facial features [11], [12]. The CNN model that was used had six layers including the convolutional, pooling and flattening layers. The CK+ [24] dataset is used for training and testing the implemented model. The CK+ dataset is a video-based action unit coded face database which involves posed expressions. It consists of images of 123 subjects.

Under the preparation and pre-processing of the dataset, the images were categorized into the seven emotional categories that are considered. Image pre-processing techniques such as grayscale, erosion, median filter and normalization were applied for input images to increase the quality of the images. The Haar-Cascade algorithm was used to segment the faces, which were then cropped and then input to the model for training and testing.

3.2 Emotion Recognition using Voice variations

Emotion recognition using voice variations is a sequence classification problem, where the input is the variant length of the acoustic signal and the output is the actual emotion. The attention mechanism based Convolutional Recurrent Neural Network (ACRNN) [25] was used for the classification, as it has been successfully applied for speech emotion recognition. It operates by producing affective-salient features by extracting the log-Mels with deltas and delta-deltas features (3-D log-Mels) from an audio file, since the deltas and delta-deltas are considered to capture the process of emotional change. Initially, a 3-D CNN was trained on the log-Mel values. Afterwards, the sequence features of 3-D CNN were fed into a Bi-directional Long Short-Term Memory (BLSTM) for temporal summarization, which identifies the most discriminative features for the final emotion expression. Since all frame-level CRNN features do not contribute equally to the representation of the emotions contained in voice, the attention layer helps to clearly identify most relevant parts and create discriminative utterance-level representations for emotion recognition.

Finally, the utterance-level representations were passed into a fully connected layer with 64 output neurons to obtain higher-level representations that enables the soft-max function to classify the utterance representations into seven numbers of emotional states. To accelerate the training process and improve performance, batch normalization was applied to the fully connected layer. Speaker-independent SER experiments on the Interactive Emotional Dyadic Motion Capture database (IEMOCAP) [26], which contains a total of 10039 utterances of female and male audio files were used to train and test the model.

3.3 Emotion Recognition using Social Media content and Speech Content

Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral [27], [28]. In our approach for the textual emotion recognition, we employed Bayesian learning to recognize and qualitify emotions in text, which is widely applied in text classification [27], [28]. The core component of this model is a multinomial Bayesian classifier which classifies the extracted emotions into seven basic emotion classes. The dataset was created from the posts, which have been shared on social media, combined with publicly available multiple sentiment analysis related datasets. Altogether the dataset consists of 94473 rows of data, which is assigned in to different emotion classes.

The implemented model performs the feature selection, classification and result conferment in a single pipeline. Fig. 1 represents the workflow that was used in the training process.

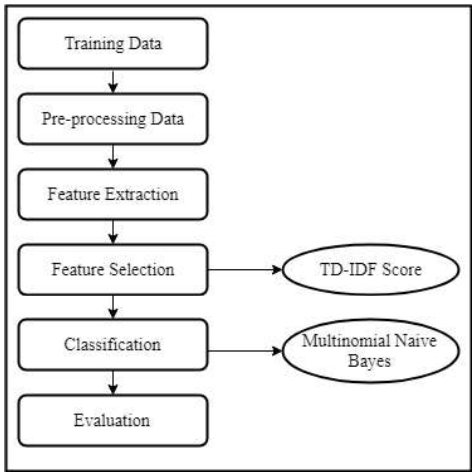

Fig. 1. Sentiment Analysis Workflow

After extracting features from the dataset, the most relevant features for classification were identified through Term Frequency-Inverse Document Frequency (TF-IDF) [29] score. TF-IDF is used for stop-words filtering in various subject fields including text summarization and classification. Multinomial Naïve Bayes considers the counts of multiple features occurred as the input features, which makes it suitable for texual emotion recognition. Both the social media posts of users as well as their speech content, which were converted from voice to text using a standard speech-to-text conversion service, were input to the model to obtain the emotion level of the user.

3.4 Ensemble Method for Aggregating the Multi-modal Emotion Predictions

In aggregating the predictions of each modality on the emotional level, it is important to capture the individual differences in expressing emotions. Each individual tends to express their emotional state in varying degrees using each modality. Thus, a better ensemble method should capture this heterogeneity in how each user expresses their emotional state. Existing work in suggest using fusion techniques such as feature level fusion, Matching-score level fusion and decision level fusion [3], [8]. Since there are three modalities to consider, the ensemble method that we introduce extends from the decision level fusion method.

In decision level fusion, features of the modalities are classified specifically by specific classifiers, and the outputs are integrated by an integration criterion. Since our focus here is to capture the individual differences in expressing emotions, we employed a user-specific learning process. This process uses a time-varying weight-driven ensemble technique, where different weights are learnt for each user for each individual modality of recognizing the emotional state. Employing this user-specific learning process to continuously adjust the weights to aggregate the emotions recognized by each modality is the novelty of the proposed approach. Eq. 1 formally expresses the proposed ensemble method.

$$P_u = \sum_{i=1}^{m} w_{u,i} p_{u,i} / \sum_{i=1}^{m} w_{u,i} \quad \textbf{(1)}$$

Here, P is the aggregated prediction of the emotional state of a user u and $p_{u,i}$ is the prediction of the emotion state of user u obtained using each modality, i. The emotion prediction obtained using each modality for each user is weighted with the respective weight $w_{u,i}$. Then the cumulative emotion prediction is normalized by dividing it by the sum of weights. The number of modalities is represented by m, and in this work the number of modalities is three, namely the facial expressions, voice variations and the speech and social media content. The weights for each modality would be user specific, capturing the heterogeneity of each user in expressing their emotions. The weights are considered to be inversely proportional to the error of predictions obtained using each modality.

$$w_{u,i} \alpha \ 1/e_{u,i} \quad \textbf{(2)}$$

Here, $e_{u,i}$ would be the error of each modality i for each user u. Since the error is user specific, the corresponding weight too will be user specific.
It should be noted that, in order to learn the user-specific modality weights, a separate user-specific learning process and a user-specific learning phase has to be employed under this method.

3.5 Evaluation

In order to test and validate the proposed ensemble method, we developed a sample client-server application, to be distributed among a sample user group and to gather the test data. Fig. 2 shows the architecture of the test application, which consists of two main modules: the user interaction module and system service module.

The mobile client is used to periodically capture the facial expressions, voice expression, social media and speech content of the users. A sample group of three voluntary users were requested to use the mobile client application over one hundred attempts to express their emotional status by speaking into the application, with their faces being visible to the camera. The mobile application would capture their facial expressions and voice that is used for the models that predict the emotional state using the facial expressions, voice variations and the speech content respectively. Further, the mobile application was granted access to their social media posts where the content of the posts were considered as another input to the text based emotion recognition model, in addition to the speech content input. The obtained data were then transferred to the server to run the predictive models and to do the aggregation. After each classification, the raw data were automatically discarded to preserve the privacy of the volunteers.

At the end of each usage, the users were requested to manually indicate their 'actual' emotional state to the system, according to their own opinion, using a scale from 0 to 10 for each of the seven emotion types. This was considered as the ground truth and the user-specific error of the predicted emotion of each modality was calculated, in comparison to this ground truth. The user-specific error of each modality was used to compute the user-specific weight of each modality, in aggregating the emotion predictions of all three models. It should be noted that the user-specific weights were automatically and continuously updated for each user, during the user-specific learning process. The obtained results for the proposed ensemble method was then compared with the standard ensemble methods such as XGBoost and AdaBoost, in which the weight of each model is negatively proportional to the general error rate of each model and not a user-specific error. Therefore, those standard ensemble methods do not require a user-specific training process or a user-specific training phase and with those methods, the user data were only used for testing their accuracy.

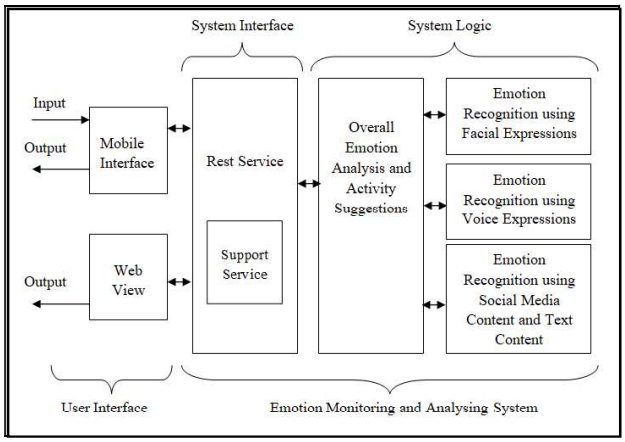

Fig. 2. System Architecture

4 Results

This section provides the results of the experiments conducted to test the accuracy of the proposed ensemble method for emotion prediction. Table 1 depicts the accuracy of the individual models that were tested using the test data that were obtained under each modality. Note that test data used for this purpose is not the test data that's obtained using the user-specific training process. Instead, the test data used for each individual model was a proportion of 20% of each data set that were used for training each model.

Table 10. Accuracy of the Models

Model Name	Accuracy
Emotion recognition using Facial Expressions	87.5%
Emotion Recognition using Voice Expressions	77.04%
Emotion Recognition using Text and Social media content	84%

Fig. 3 shows the accuracy of the predicted emotional states of the three sample users by comparing the accuracy the emotional state predicted using each individual modality, aggregation of two modalities and the aggregation of all three modalities considered. Here, FER refers to Facial expression based emotion recognition modality, SER denotes the Speech or voice variation based emotion recognition modality and TER is the Text based emotion recognition modality, respectively. According to the results, it is evident that the error rate of each modality varies for each user, further supporting the fact that each user may express their emotions in varying degrees using each modality. Further, the results show that in general, when all the modalities are used, the proposed ensemble method gives better predictions of the user's current emotional state.

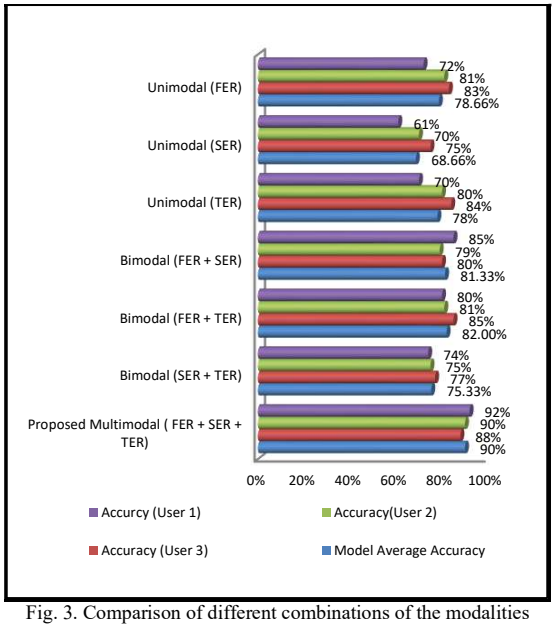

Fig. 3. Comparison of different combinations of the modalities

Table 2 shows the comparison of the proposed ensemble method with the standard ensemble methods such as XGboost and AdaBoost ensemble methods. Based on the comparison, it is evident that the proposed ensemble method outperforms the other ensemble methods considered. This may be due to the fact that the proposed ensemble method considers the user-specific nature of the modalities while the other standard ensemble method have a generic weight that are applied uniformly to each user.

Table 2. Comparison between different Ensemble Methods

	Accuracy
XGBoost	86%
AdaBoost	85%
Proposed ensemble method	92%

5 Discussion and Future works

In this study, we have presented a novel ensemble method to recognize the emotional state of an individual, using a multimodal approach. It captures the heterogeneity of users in how they express their emotions using different modalities such as facial expressions and voice variations. The results indicate that the proposed ensemble method is more accurate in recognizing the user emotions,, compared to the standard ensemble methods that do not consider the heterogeneity of the users in expressing their emotional state. The results further suggest that the three modalities of facial expressions, voice expressions and speech and social media content are combined, the emotion recognition is more accurate, in comparison to the scenario when only one or two of the modalities are used to predict the emotional state.

There are several limitations of the proposed approach. Firstly, the number of users in the test group can be increased to further validate the results. Also, though we considered only three modalities in this experiment, additional modalities such bodily postures, travel patterns can be incorporated to the proposed ensemble method, provided that the relevant data sources are available. Also, it should be noted that the proposed ensemble method would require a separate user-specific learning process unlike the standard ensemble methods such as AdaBoost. This can be regarded as a limitation of the proposed method. Further, in real-world applications such recommendation systems or mental healthcare applications that may use automatically recognized emotional state of a user as an input, the user-specific training phase has to precede the actual usage of the application.

While we only considered user-specific weights for the modalities, it may be possible to cluster the users based on the demographic details of the users and assign an averaged set of weights for the users within the same cluster. This may be particularly useful if it's necessary to avoid a separate user-specific training process or the cluster-specific weighs may be used as a particular user's initial weights if the user has not given any prior feedback on the actual emotional state. However, it is necessary to have sizable dataset of the users, their modality weights and demographic details, in order to apply a cluster-specific set of modality weights.

Even though we employed a greedy and linear method to update the modality weights at the user-learning stage, a non-linear and non-greedy learning technique such as Deep Reinforcement learning may be used to update the modality weights, which may lead to more accurate aggregated predictions of a user's emotional state.

Although this work mainly focuses on emotion recognition, the proposed ensemble method may be applicable for other predictive problems in the domain of human-computer interaction, such as behavior prediction, where an aggregation of multiple modalities may be applicable and the user-specific variations in those modalities need to be preserved.

References

1. Emotions, www.thesaurus.com/browse/emotions (2018)
2. Bevilacqua, V., D'Ambruoso, D., Mandolino, G., Marco Suma.: A new tool to support diagnosis of neurological disorders by means of facial expressions. In: IEEE International Symposium on Medical Measurements and Applications (2011)
3. Busso, C. et al,: Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proceedings of the 6th international conference on Multimodal interfaces, pp. 205-211 (2004)
4. Ekman, P., and Friesen, W.V.: Constants across cultures in the face and emotion. J. Pers. Soc. Psychol., vol. 17, no. 2, pp. 124–129 (1971)
5. Whelton, W.J., Emotional processes in psychotherapy: Evidence across therapeutic modalities.: Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice, 11(1), pp.58-71 (2004)
6. Bänziger, T., Mortillaro, M., and Scherer, K.R.: Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. Emotion, 12(5), p.1161 (2012)
7. Vinola, C., and Vimaladevi, K.: A Survey on Human Emotion Recognition Approaches, Databases and Applications. ELCVIA Electronic Letters on Computer Vision and Image Analysis, vol. 14, no. 2, p. 24 (2015)
8. Kumar, K.M.A., Kiran, B.R., Sriyas, B.R., Victor, S.J.: A Multimodal Approach To Detect Users Emotion. Procedia Computer Science, vol. 70, pp. 296–303 (2015)
9. Meftah, I.T., Thanh, N.L., Amar, C.B.: Multimodal Approach for Emotion Recognition Using a Formal Computational Model.: International Journal of Applied Evolutionary Computation, vol. 4, no. 3, pp. 11–25 (2013)
10. Cid, F., Manso, L.J., Núñez, P.: A novel multimodal emotion recognition approach for affective human robot interaction. In: CEUR Workshop Proc., vol. 1540 (2015)
11. Alizadeh., S, Fazel, A.: Convolutional Neural Networks for Facial Expression Recognition.: In arXiv preprint (2017)
12. Shan, K., Guo, J., You, W., Lu, D., Bie, R.: Automatic Facial Expression Recognition Based on a Deep Convolutional-Neural-Network Structure. IEEE Comput. Soc., pp. 123–128 (2017)
13. Lopes, A.T., Aguiar, E. de, De Souza, A.F., Oliveira-Santos. T.: Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order. Pattern Recognit., vol. 61, pp. 610–628 (2017)
14. Badshah, A.M., Ahmad, J., Rahim, N., Baik, S. W.: Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. In: Int. Conf. Platf. Technol. Serv. (2017)
15. Fran E.I., Ispas, I., Dragomir, V., Dasc, E.M., Zoltan, A., Stoica, I.C.: Voice Based Emotion Recognition with Convolutional Neural Networks for Companion Robots. Rom. J. Inf. Sci. Technol. (2017)

16. Alghifari, M.F., Gunawan, T.S., Kartiwi, M.: Speech emotion recognition using deep feedforward neural network. Indones. J. Electr. Eng. Comput. Sci. (2018)
17. Chaffar, S., Inkpen, D.: Using a heterogeneous dataset for emotion analysis in text. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6657 LNAI, pp. 62–67 (2011)
18. Jain, U., Sandhu. A.: A Review on the Emotion Detection from Text using Machine Learning Technique. Emotion. (2009)
19. Perikos, I., Hatzilygeroudis, I.: Recognizing emotions in text using ensemble of classifiers. Engineering Applications of Artificial Intelligence (2016)
20. Morrison, D., Wang, R., Silva, L.C.D.: Ensemble methods for spoken emotion recognition in call-centres. Institute of Information Sciences and Technology, New Zealand, pp. 98–112 (2007)
21. Lingenfelser, F., Wagner, J., André, E.: A Systematic Discussion of Fusion Techniques for Multi-Modal Affect Recognition Tasks. In: Proceedings of the 13th international conference on multimodal interfaces, Alicante, Spain, pp. 19-26 (2011)
22. Padgett, C., Cottrell, G.: Representing Face Images for Emotion Classification. Advances in Neural Information Processing Systems, Vol. 9, MIT Press, Cambridge, MA, pp. 894 –900 (1997)
23. Mahto, S., Yadav, Y.: A Survey on Various Facial Expression Recognition Techniques. Ijareeie, vol. 3, no. 11, pp. 1–9 (2014)
24. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The Extended Cohn-Kanade Dataset (CK): A complete dataset for action unit and emotion-specified expression. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition – Workshops (2010
25. Chen, M., He, X., Yang, J., Zhang. H.: 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Processing Letters, 25(10), pp.1440-1444 (2018)
26. Busso, C., et al.: IEMOCAP: interactive emotional dyadic motion capture database. Language Resources & Evaluation 42.4:335 (2008)
27. Zucco, C, Calabrese, B., Cannataro, M.: Sentiment analysis and affective computing for depression monitoring. In: IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2017
28. Kibriya, A.K., Frank, E., Pfahringer, B., Holmes, G.: Multinomial naive bayes for text categorization revisited. In: Proceedings of Australasian Joint Conference on Artificial Intelligence, Springer, Berlin, Heidelberg, pp. 488-499 (2004)
29. Haddi, E., Shi, X.L., Y.: The role of text pre-processing in sentiment analysis. Procedia Computer Science, 17, pp.26-32 (2013)