Journal of Information Systems and Telecommunication
Vol.10, No.2, April-June 2022, 89-101
http://jist.acecr.org
ISSN 2322-1437 / ESSN:2345-2773

Hierarchical Weighted Framework for Emotional Distress Detection using Personalized Affective Cues

Nagesh N Jadhay[1*], Rekha Sugandhi[2]

[1].Department of Computer science and Engineering, MIT School of Engineering, MIT ADT University, India
[2].Department of Information Technology, MIT School of Engineering, MIT ADT University, India

Received: 09 Jul 2021 / Revised: 04 Oct 2021 / Accepted: 08 Nov 2021
DOI:

Abstract

Emotional distress detection has become a hot topic of research in recent years due to concerns related to mental health and complex nature distress identification. One of the challenging tasks is to use non-invasive technology to understand and detect emotional distress in humans. Personalized affective cues provide a non-invasive approach considering visual, vocal, and verbal cues to recognize the affective state. In this paper, we are proposing a multimodal hierarchical weighted framework to recognize emotional distress. We are utilizing negative emotions to detect the unapparent behavior of the person. To capture facial cues, we have employed hybrid models consisting of a transfer learned residual network and CNN models. Extracted facial cue features are processed and fused at decision using a weighted approach. For audio cues, we employed two different models exploiting the LSTM and CNN capabilities fusing the results at the decision level. For textual cues, we used a BERT transformer to learn extracted features. We have proposed a novel decision level adaptive hierarchical weighted algorithm to fuse the results of the different modalities. The proposed algorithm has been used to detect the emotional distress of a person. Hence, we have proposed a novel algorithm for the detection of emotional distress based on visual, verbal, and vocal cues. Experiments on multiple datasets like FER2013, JAFFE, CK+, RAVDESS, TESS, ISEAR, Emotion Stimulus dataset, and Daily-Dialog dataset demonstrates the effectiveness and usability of the proposed architecture. Experiments on the enterface'05 dataset for distress detection has demonstrated significant results.

Keywords: Convolution Neural Network; Long Short-Term Memory; Transformers; Hierarchical Fusion; Distress Detection.

1- Introduction

In the current scenario of a global pandemic, every individual is fighting his or her own battles on financial, social, and medical fronts. The current situation is making humans undergo emotional fluctuations, resulting in challenges in mental health. Paul Ekman proposed seven basic emotions like happy, sad, fear, angry, disgust, surprise and neutral [1]. Every person undergoes various emotions based on the context he or she is in. It is natural for a human to exhibit different emotions, however, the prevalent appearance of some emotions like sadness, fear and anger can be alarming. For example, consider a small video from Youtube of 10 seconds with 40 frames and 10 frames per second frame rate, the total number of emotional changes that happened are around 40. The term 'distress' is related to the affective state that arises in vicious or furious situations [2]. The National Comprehensive Cancer Network has defined distress. They quote 'Distress is a multifactorial unpleasant emotional experience of a psychological (cognitive, behavioral, emotional), social, and/or spiritual nature. Distress encompasses a range of common feelings of vulnerability, sadness, and fears that can cause depression, anxiety, panic, social isolation, and existential and spiritual crisis' [3]. Unanticipated external events are not only the cause of distress but also internal states like thoughts, chronic behaviors, and feelings of the person. Depression and anxiety disorders follow the prevalent distressed condition.

The affective computing community is working closely with psychologists in the identification, prevention, and treatment of mental disorders. Recent studies signify the use of complex multi-modal systems preferred over single modal architectures. Humans express and communicate their emotions in a multimodal way. Affective cues such as visual, auditory, and textual are exploited parallelly and cognitively to extract affective and profound information communicated during the interaction. This information can be used to identify emotional distress if analyzed and used

✉ **Nagesh N Jadhay**
nagesh.jadhav@mituniversity.edu.in

effectively. Mehrabian and Ferris, in their research, revealed the contribution of audio, visual and textual cues during the communication. According to the authors, 55% of communication is visual, 38% is vocal and 7% is verbal [4]. In recent years, people are inclined towards social media, which adds to a large amount of video, audio, and text data generation. Social media platforms are also used to express opinions about products, services, or people. Information can be analyzed to extract required details to identify the emotional states of the person. The multi-modal frameworks are still a challenge in computer science due to two main reasons. 1) It is difficult to extract useful features from the audio, video, and text data. 2) The different dimensions of each modality make it difficult to fuse the feature and process it.

In this paper, we are proposing a hierarchical weighted framework for the detection of emotional distress based on audio, video, and text modalities. Most of the literature focuses on multi-modal emotion recognition using video and audio cues. We have also considered textual cues to understand the affective state of the person. Considering the complexities and challenges in facial emotion detection we have employed a hybrid approach on multiple face emotion datasets. We have used two native convolutional neural networks (CNN) approaches and one transfer learning approach to achieve better results. The purpose behind using multiple datasets is to have a wide variety of facial expressions from people across the globe. We have also developed our dataset of face emotions expressions, where all the images are taken in wild with an available resource like a smartphone. Over 43 students from the computer science and engineering department volunteered for this activity. For audio affective cues, we used both convolutional neural networks and Long Short-Term Memory (LSTM) architectures. LSTM performed reasonably well in comparison to CNN. The textual cues are also very challenging to handle. Wrong interpretation of context will lead to the wrong semantic of the sentence. Initially, we worked with bi-directional LSTM, however, we found Bidirectional Encoder Representations from Transformers (BERT) are very effective and precise for textual cue processing. Fusing the results from the different modalities is a challenging problem. To fuse the results, we have proposed a novel hierarchical weighted framework. To detect the emotional distress of a person we have also proposed and implemented the algorithm based on the negative emotions demonstrated by the person.

2- Related Work

Emotional distress has become one of the common mental illnesses if goes unattended may lead to anxiety, depression moreover into suicidal tendencies. Prolonged stress also results in sleeplessness, mood swings and lack of attention. Emotional distress is directly related to mental disorders. The clinical approaches in diagnosing the distress include manual intervention and questionaries to answer. Automating this process with optimized deep learning approaches would be groundbreaking and it will help medical and psychological practitioners significantly.

In past, affective state recognition including facial cues and audio cues been studied by many researchers. Most of the work focuses on the utilization of audio and video cues ignoring textual affective cues. Thomas et al. [5] have used audio and visual information to understand the affects. The research focuses on identifying valence and arousal for the affects. Dataset used is the MediaEval 2015 dataset. Both audio and video representations are fused and trained using an SVM classifier. Each of the features is also trained independently and used an ensemble approach to fuse the results. The combining video and audio features have demonstrated good results compared to handling a single modality. Affective internet of things is an active area of research that can detect the affective state of the human. Wearable systems which are the combination of recent market technology and smart sensors are an important part of affective computing. Miranda et al. have used Galvin skin resistance and blood volume pulse as modalities to detect fear emotion, which falls in the negative quadrant of the valence and arousal. Using wearable computing for affective state recognition could be an expensive affair when it is considered for real-time use and applications [6]. The authors [7] have proposed a multi-modal fusion scheme of audio-video features to detect the depression of the person. The extracted audio and video features from the stream of data, initially processed using principal component analysis to reduce the redundancy, which is further provided as input to the epsilon support vector regression model for the prediction. The predicted values are finally combined with the local linear regression model to predict the result. For visual cues divergence curl shear descriptors, space temporal interesting points and head pose features are used as the features to detect depression. For the audio motion history histogram, a bag of words and vector of locally aggregated descriptors are used to examine the audio of the person. The proposed method uses the 3DCLS hybrid model which is the combination of 3D CNN and convolutional LSTM. The text modality is processed using the CNN-RNN hybrid model. CNN is used to extract features from the text while RNN is used to predict the emotion. For the visual data, C3D is used to extract Spatio-temporal features and CNN LSTM predicts the feature sequences. Features from the audio data are extracted using the OpenSmile tool and processed using SVM [8]. Zeinab and Saeed proposed the MoBEL model [9], which is a combination of expert neural networks and brain-inspired learning algorithms. It works in two stages. In the first

stage, CNN and RNN are applied to extract high-level features while in the second stage model is trained to learn audio-video features. Luu-Ngoc Do et al. in their work used a hybrid CNN RNN model to detect the facial emotions of the person. The JAFFE and MMI are the datasets used for the experimentation. The modality considered is limited to only visual cues [10].

Jain et al. used a hybrid CNN RNN model to detect the facial emotions of the person. The JAFFE and MMI are the datasets used for the experimentation. The modality considered is limited to only visual cues [11]. Kaya et al. have proposed a multimodal approach where the input is processed to extract faces from the video. The extracted faces are trained using a pre-trained VGG16 model. The audio cue is processed separately, and results are fused at the decision level. For classification model learning, kernel extreme learning machine and partial least square regression are used. Finally, the results are combined using a weighted fusion of scores [12]. Zhang et al. discuss the multi-modal approach for speech emotion recognition. The proposed method creates three audio inputs for the CNN model. The First 1D CNN model uses raw waveform, 2D CNN uses MFCC feature while 3D CNN exploits temporal-spatial features of the input. The score level fusion is performed to attain the final prediction [13]. Yan et al. have defined a multi cue fusion emotion recognition framework based on the audio signal, facial texture, and facial landmarks. To capture change in facial texture the cascaded convolutional neural network and Bi-directional RNN is used. Audio features are extracted using CNN and stored in a matrix for further processing. The faces are trained using a pre-trained VGG16 network. The RNN model is employed on extracted facial texture feature to recognize dynamic differences in facial feature sequences. For audio, OpenSmile is used to extract the features, further CNN is used for classification. Both feature level fusion and decision levels strategies are exploited to achieve results [14]. Bendjoudi et al. proposed architecture for the visual cues. An input image is processed by a scene detector module. Body module is used to process cropped images. VGG16 pre-trained network and Xception network is used for scene detector and body detector module, respectively. Authors have proposed the multi-label focal loss function to calculate the loss on the Emotic dataset [15]. Hao et al. proposed an emotion recognition method based on multi-task and ensemble learning using audio-video features. Both the deep features and manual features of audio-video cues are extracted and presented to different algorithms for processing. The proposed architecture consists of four models, CNN for mel spectrogram, SVM from Interspeech2010 features, CNN for facial emotion recognition and SVM for LBP features. A blending algorithm is used to fuse the results from different classifiers [16]. Tzirakis et al. have considered visual and auditory cues for emotional recognition. CNN is used to extract features from speech while pre-trained ResNet50 is used to extract visual features. The RECOLA dataset has been used to test the developed model. The output from both modalities are combined in a feature vector and passed to LSTM for the prediction of the result [17].

Majumder et al. have worked on three modalities like audio, video, and text. In the initial phase, features are extracted individually from each of the modalities. The proposed hierarchical approach starts with a bimodal fusion of modalities followed by trimodal fusion. The experiments were performed on the IEMOCAP and CMU-MOSI databases [18]. The proposed method in [19] considers three modalities, audio, video and text for data analysis. The experiments were performed on e'NTERFACE database. Each modality is trained individually on a different dataset. For visual, auditory and text-based cues, CK+ and ISEAR databases are used respectively. Yaxiong Ma et al. proposed a deep fusion method for audio and visual cues. To remove cross pollution in audio data and redundancy in visual cues, cross-modal noise modelling has been used. Audio feature extraction is done using 2D CNN while video feature extraction is done using 3D CNN. Finally, deep belief networks are employed for the nonlinear fusion of extracted features [20]. Guo and the team worked on video and audio data for video content analysis. VGG16 pre-trained is used to extract the video features and audio features are extracted using the OpenSMILE tool. Following end-to-end training, two subtasks are introduced i.e., classification and regression. The classification task focuses on classifying fear-induced videos and the regression task predicted the arousal and valence values of the user [21]. Poria et al. used audio, video, and text data to perform sentiment analysis of the video. Both feature level fusion and decision level fusion are utilized to merge different modalities [22]. Noroozi et al. proposed architecture to work on the audio and video data. For audio data features like MFCC, prosodic and filter bank energies are extracted and for visual cues, CNN is applied to extract the required features. Finally, the confidence outputs of multiple classifiers from the different modalities are fused to fetch results [23]. In [24] Audiovisual cues are considered to learn affective features using a combination CNN and 3D CNN model. The extracted audio-video features are fused using the DBN network. Li and Liu [25] have developed 1D CNN network and multi-layer perceptron network to detect stress using physiological signals. The input has been classified into baseline, stressed and amused states. Bobade and Vani have worked on multimodal physiological signals for the detection of stress using machine learning algorithms. Results in the paper demonstrated deep neural networks performs better compared to conventional machine learning algorithms [26]. Zhang et al. have discussed video based stress

detection. The authors have used CNN for the implementation. The results of facial expression and facial units are combined using weighted integration of local and global attention [27].

The literature survey has identified the following research gaps, which are addressed in this paper.

- Most of the multi-modal approaches focus on physiological signals.
- Results of training and testing are demonstrated on the same dataset. Cross data result validation is missing in many approaches
- The early feature fusion approach is used by most researchers which raises the requirement of high computational resources for model training.
- We found limited literature related to stress detection using deep learning and late fusion approaches.

3- Proposed Work

The overview of the research methodology is shown in Fig. 1. Where audio-video input has been given to the system for processing. In the initial step video frames, audio, and text data has been separated from the input and assigned to subsequent block for further feature extraction and processing. The working of each block has been discussed in detail in the following sections.

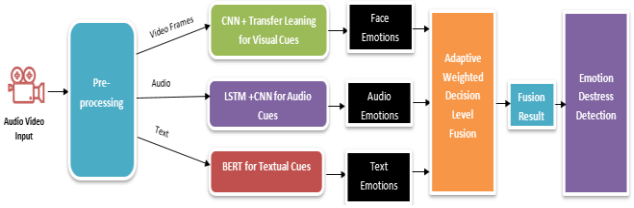

Fig. 1 Proposed Work

3-1- Visual Cues

Most of the literature has identified that the facial cues are very challenging to work due to unavailability of the data, variation facial structures, ethnicity of the person, wrong perception of emotion by annotators of dataset etc. Many datasets also have a biased distribution of emotions in the dataset. To overcome these issues, we have worked with three different datasets like Face Emotion Recognition 2013 (FER 2013), Japanese Female Facial Expressions (JAFFE) and Cohn Kanade plus (CK+) datasets [28][29][30]. A detailed overview of facial cues processing and prediction is shown in Fig. 2.

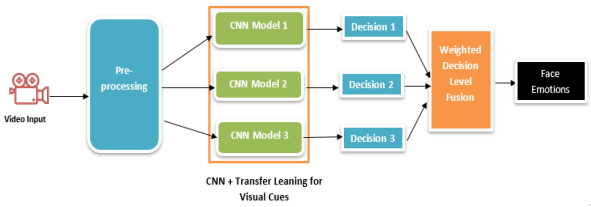

Fig. 2 Visual cue processing for facial emotion recognition

3-1-1- Preprocessing

The video data received is in different formats and dimensions. The preprocessing is performed on the video to extract the faces from the video frames. This is done using two steps:

1. Face Detection: Many face detection algorithms are available to detect and extract faces from video frames. We have worked with two approaches classical Voila-Jones algorithm and multitask cascaded convolutional networks. Both algorithms worked well for the input provided. Considering the amount of time and memory required to process the input, we decided to use a lightweight Voila-Jones algorithm.
2. Normalization: The detected faces are resized and normalized as per the requirement of CNN models. Each detected face has been resized to (197,197,3), (128,128,3), and (48,48,3) respectively.

3-1-2- Datasets

For face emotion recognition we have used three datasets, FER2013, JAFFE and CK+. The FER2013 is an extensive dataset with 35887 grey images of dimension (48*48*1) depicting 7 basic emotions. The distribution of classes is given as Angry:4593, Disgust:547, Fear:5121, Happy:8989, Sad:6077, Surprise:4002, and Neutral:6198 images. We can see from the distribution that the dataset has a small number of images depicting ‘disgust’ emotion. The JAFFE dataset has 213 images of Japanese females portraying 7 basic images. All the images are of dimension 256*256*1. The third dataset we have used for model training is the CK+ dataset. It is a very popular dataset consisting of 593 image sequences. Apart from basic emotion, it also has images representing ‘contempt’ emotion. The sample pictures from the database are shown in Fig. 3.

Fig. 3 Facial Cue Datasets

3-1-3- CNN Model I

For the first CNN model, we have adopted already trained ResNet50 [31] and VGGFace model, instead of developing the model from the scratch. ResNet50 is one of the popular models used in computer vision. It consists of 48 convolution layers, one max pooling and one average pooling layer. The ResNet50 was developed to overcome the problem of vanishing gradient and results in saturation as the model goes deep in architecture. ResNet50 was trained on the Imagenet database and can classify 1000 objects. The key part of ResNet50 is identify block which takes residual connections. The Working of identify block is shown in Fig. 4.

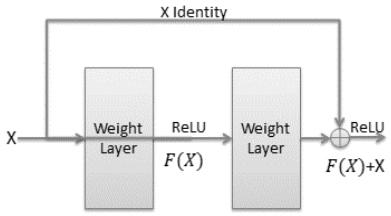

Fig. 4 Residual Identity Block

The Resnet50 has achieved great success in recognizing required features from the images automatically. VGGFace model has performed greatly in recognizing the faces. Therefore, the already trained ResNet50 model along with the weight of VGGFace can be easily transferred to learn the feature of the images from the FER2013 dataset. However, the images in the FER2013 dataset are of size 48*48*1. All these images are required to resize to the dimension that suits ResNet50 and VGGFace models. Each image has been resized to dimension (197,197,3). To avoid overfitting, we employed image augmentation to generate runtime tensor image data. The top layers of the model are fully trained by freezing convolutional layers. After top layer training, the convolutional layers of ResNet50 are fine-tuned to achieve high accuracy for both validation and test set.

3-1-4- CNN Model II

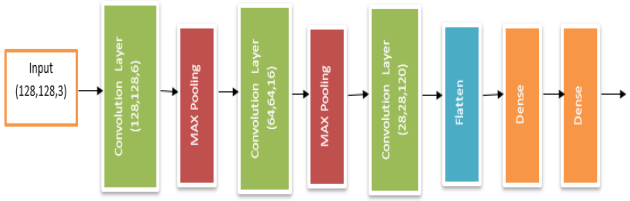

Fig. 5 Architecture for JAFFE dataset

The second model is designed for the JAFFE database. The FER2013 dataset has very limited images displaying 'disgust' emotion. To avoid misclassification of disgust emotion we have incorporated the JAFFE dataset. The CNN model is developed and fine-tuned to work on the JAFFE dataset. The basic architecture is shown in Fig. 5. The input of dimension (128,128,3) has passed to the CNN architecture. The model consists of three convolutional layers, max-pooling layer is interspersed between the convolutional layers to downsample the feature map. The flatten is layer followed by two dense layers and softmax activation is used as the output unit. The softmax activation is denoted by the following expression,

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{i=1}^{N} e^{z_i}} \quad (1)$$

$$z_i = weight \times input^T$$

3-1-5- CNN Model III

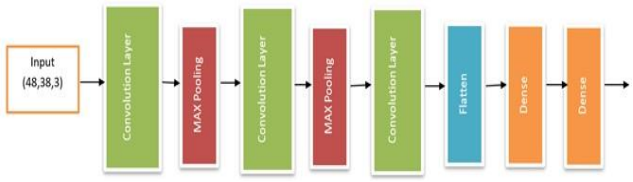

Fig. 6: Architecture for CK+ dataset

To improve the overall accuracy of facial emotion recognition, we have considered another CNN model on the CK+ dataset. All images are resized to (48,48,3) dimensions. To split the data, k-fold validation has been used with a value for a number of splits is equal to 5. The model has shown significant accuracy in the training and testing phase. The model architecture with dimensions is displayed in Fig. 6. All the models produce their own decision upon processing the various dimension inputs. All decisions are fused using a weighted approach.

3-1-6- Visual Cue Emotion Recognition

Emotions = {Anger, Disgust, Fear, Happy, Sad, Surprise, Neutral}
TA_video = {w1, w2, w3}
Where TA_video is normalized test accuracy of video modality for CNN model 1, model 2 and model 3 respectively. The weight normalization is given as,

$$w_i = \frac{M_{acc}}{Acc_{Total}} \quad (2)$$

Where M_acc is the test case accuracy of the model.
Steps for the emotion recognition from video are given below
Step 1: Video input
Step 2: Pre-process the input video to extract the faces from video frames and resize the frames as per required dimensions.
Step 3: Classification – Apply all models concurrently to classify the video frame in one of the classes from Emotions.
Step 4: Apply the voting approach to select the final facial emotion
Step 5: In the case of 3 indifferent decisions, apply weights w1, w2, and w3 respectively to each model's decision.
Step 6: Make final prediction using following equations,

$$face_{emotion} = argmax\left\{\sum_{1}^{L} argmax\left(w1 \times Decision_1, w2 \times Decision_2, w3 \times Decision_3\right)\right\} \quad (3)$$

Where $Decision_1$, $Decision_2$, *and* $Decision_3$ are the probability values associated with each decision, L is the length of the video. Based on the emotion of each frame detected, higher count emotion is selected as the final $face_{emotion}$ of the video, ignoring neutral emotion.

3-2- Audio Cues

Many times, a person fakes their facial expressions to hide the actual emotions he is undergoing. So, relying on visual cues would add inconsistencies in the recognition of emotional distress. Audio cues, when combined with visual cues provide significant results. Different audio features describe different emotions. For example, anger emotion can be defined using pitch frequency, rapid speech rate and high energy. Similarly, sadness can be described as low energy and pitch frequency. Literature suggests Mel Frequency Cepstral Coefficients (MFCC), spectral energy distribution, the intensity of speech, pitch, Zero Crossing Density (ZCD) are the important and widely used features to recognize emotions from audio signals [32]. Mel Frequency Cepstral Coefficient: Mel scale is related to the perceived frequency or pitch of actual measured frequency. The Mel scale is represented using the following formula,

$$M(f) = 1125 \times \ln\left(1 + \frac{f}{700}\right) \quad (4)$$

Given the above equation, Mel frequency is represented as,

$$M^{-1}(m) = 700 \times \left(exp\left(\frac{m}{1125}\right) - 1\right) \quad (5)$$

The Cepstral Coefficients (CC) can be exploited to split the original signal from the filter. Spectral details of a signal can be extracted by truncating the signal at different frequencies. To calculate Cepstrum, Discrete Fourier Transform (DFT) of the log magnitude of the DFT of the signal is calculated. The model proposed by Davis et al. [33] to calculate MFCC is given in equation (6),

$$MFCC_i = \sum_{\theta}^{N} \cos\left(i(\theta - 1)\frac{\pi}{N}\right),\ i = 1, 2, 3, \ldots N \quad (6)$$

Where M is cepstral coefficients, Theta is long energy output and N is a number of triangular bandpass filters.
For audio cues, we developed two models, the first model exploiting the capabilities of CNN and the second model using the LSTM approach. Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [34] and Toronto emotional speech set (TESS) [35] datasets of emotional audio speech are used. The proposed architecture for audio cue recognition is shown in Fig. 7.

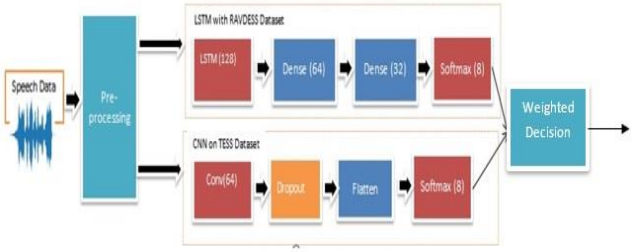

Fig. 7 Architecture for Audio cues emotion recognition

We employed the CNN model on the TESS dataset and the LSTM model on RAVDESS dataset. We used 40 MFCC features in both models for preprocessing. Audio data has been extracted from video and preprocessed to fetch MFCC features from the input. The results of both approaches are discussed in the experimentation and results section. However, the LSTM model outperformed CNN model in terms of accuracy and generalization. We fused results of both the models using a weighted approach where higher weight is assigned to the LSTM model based on the test accuracy achieved. See equation (7).

$$Audio_{decision} = argmax\left[w1_{cnn} \times P(D_1), w2_{lstm} \times P(D_2)\right] \quad (7)$$

3-2-1- Datasets

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), the dataset contains 7356 files for 24 professional actors, 12 male and 12 female. The speech

data depicts happy, calm, sad, angry, fear, surprise, and disgust emotions. The intensity of emotions portrayed is both normal and high. For the experimentation, we have used speech audio files (1440 files). Toronto emotional speech set (TESS), the dataset contains speech data, spoken by two actresses over 200 target words. It also portrays seven basic emotions.

3-3- Textual Cues

Recognizing affective states or emotions from the text is a challenging task. The reasons for the same can be listed as, word ambiguities, the complexity of meanings, writing style, different languages, different cultures and many more. Emotion detection from a text can be done using a rule-based approach, machine learning approach or hybrid approach. While recognizing emotion from the text, it is also necessary to understand the semantics and context of a sentence. We used a bi-directional LSTM and BERT transformer for the experimentation on text datasets. However, the BERT transformer performed significantly well in comparison with bi-directional LSTM [36]. We used International Survey on Emotion Antecedents and Reactions (ISEAR) dataset [37], Emotion Stimulus dataset [38] and DailyDialog dataset [39]. For the textual cues affect detection we considered only emotions, neutral, happy, sad, angry and fear.

3-4- Decision Level Hierarchical Weighted Fusion and Distress Detection

3-4-1- Decision Level Hierarchical Weighted Fusion

One of the key challenges in the multi-modal approach is to fuse the different modalities to generate a result. The majority of the literature suggests three approaches i.e., early fusion, late fusion, and hybrid fusion. A most common solution to fuse different modalities is to have early or feature level fusion. Early fusion suffers from several drawbacks. If we have a large number of features fusing them will lead to low accuracy if the training dataset is small. Format incompatibility and temporal features mapping would be challenging for early fusion. Lastly, the requirement of huge computational resources to manage high dimensional data is another significant challenge [40] [41]. In this paper, we have utilized decision level hierarchical weighted fusion. We have observed each modality has a different way of learning parameters. This results in variation in the accuracy values of each modality. So, to have consistency in decision making, we have considered the test accuracy of the datasets as a weighted parameter for further calculation. We created a hierarchy of decisions before making a final decision. The rule for the calculation of visual cue prediction probability is given in equation (8),

$$P(F_E) = argmax\left[w_1 \times P(model_{f1}), w2 \times P(model_{f2}), w_3 \times P(model_{f3})\right]$$
$$P(A_E) = argmax\left[w_1 \times P(model_{a1}),\ w_2 \times P(model_{a2})\right]$$
$$P(T_E) = argmax(\sigma(z_i)) \quad (8)$$

Where wi is normalized as per the formula given in equation (2). P(FE), P(AE) and P(TE) are the probabilities of face, audio, and text emotions, respectively. The result of all fused modalities is given as,

$$Affective_{state} = argmax\left\{W_1 \times \left[\frac{P(F_E)}{S_A}\right], W_2 \times \left[\frac{P(A_E)}{S_A}\right], W3 \times \left[\frac{P(T_E)}{S_A}\right]\right\}$$
$$where,\ S_A = P(F_E) + P(A_E) + P(T_E) \quad (9)$$

For the final decision, we have considered static weight values (W1, W2, W3) assigned to visual, vocal, and verbal cues, respectively. We have followed the Mehrabian and Ferris approach for the contribution of each cue in the communication. We tweaked the standard values mentioned and derived new weight values. For visual cues, we have considered 0.40, for audio 0.35 and text 0.2.

3-4-2- Emotional Distress Detection

Emotions are a vital part of human life. Every human feels all basic seven emotions, those may be comfortable or uncomfortable. Distress is different from emotions. We undergo different emotions regularly and we can observe many fluctuations in the emotion. The negative emotions like sadness, fear, anger, and disgust if observed consistently, it is an indication of a threatening situation. If left unattended, this emotional combust may lead to severe mental health problems. Here we are proposing the automated approach to detect emotional distress. Our system will be observing and analyzing the affective states of the person. More specifically, negative emotions like fear, sadness and anger need to be observed for their prevalent occurrences. Our proposed system will alert the person about his frequent emotional changes. The detailed flowchart is shown in Fig. 8.

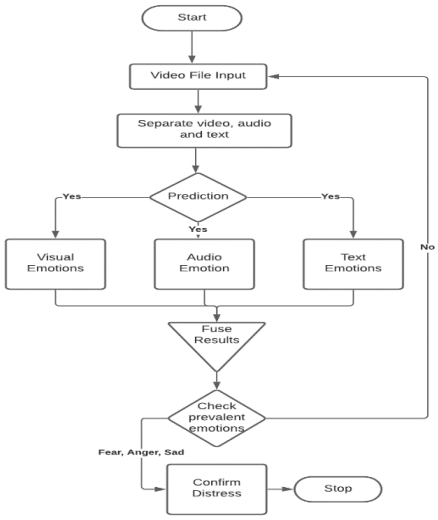

Fig. 8 Emotional Distress Detection

4- Experiments and Results

We have conducted experiments on various datasets depending on the various affective cues. Table 1 below shows different datasets we have used.

Table 1: Affective cues and datasets

Sr. No.	Affective cue/ Modality	Datasets
	Visual cues	FER2013 CK + JAFFE
	Vocal Cues / Audio data	RAVDESS TESS
	Verbal cues / Text data	ISEAR Daily Dialog Emotional Stimulus

4-1- General Settings

Given a video clip as input, data is extracted from it. Audio is extracted from video and audio is converted into text format for further processing. An entire experiment has been run into a Google collaboratory environment with 12 GB of RAM and GPU support. We have limited video length to 40 frames with frames per second rate varies from 15 fps to 25 fps.

4-2- Experiments on Visual Cues

4-2-1- Experiments with Fer2013 Dataset

FER2013 dataset is very challenging to work on because of biased distribution of samples, mislabeled emotions etc. Also, most of the models we developed, were overfitting for the FER2013 dataset. We have used transfer learning using the pre-trained ResNet50 model and weights of the VGGFace model to detect faces and recognize emotions. Every input image has been normalized to the mean of the dataset. Input has been resized to the dimension of (197,197,3). For better accuracy, we applied image augmentation with (rotation range, shear range, zoom range, horizontal flip) parameters. We have used Adam optimizer with a learning rate of 0.0001 and epsilon value with (1*e-08). With the fine-tuned model we have achieved 71.25% test accuracy. The confusion matrix and classification report are shown in Fig. 9 and Table 2, respectively.

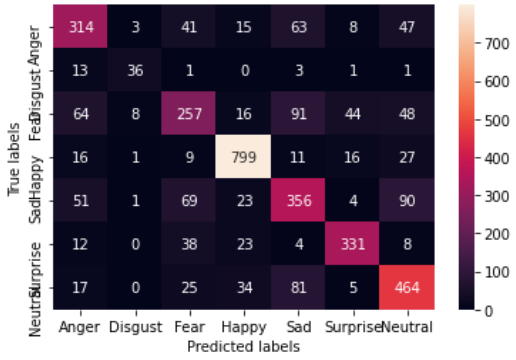

Fig. 9 CNN model 1 confusion matrix

Table 2: CNN Model 1 Classification Report

	precision	recall	f1-score	support
Anger	0.64	0.64	0.64	491
Disgust	0.73	0.65	0.69	55
Fear	0.58	0.49	0.53	528
Happy	0.88	0.91	0.89	879
Sad	0.58	0.6	0.59	594
Surprise	0.81	0.8	0.8	416
Neutral	0.68	0.74	0.71	626
Accuracy			0.71	3589
Macro Avg	0.7	0.69	0.69	3589
Weighted Avg	0.71	0.71	0.71	3589

4-2-2- Experiments with JAFEE Dataset

Anger, fear and sadness are the important emotions considered for distress detection. From the results of FER2013 dataset experiments, it is observed that accuracy for these negative emotions varies between 53% to 64%. To improve the overall performance of visual cues recognition, we have considered the JAFFE dataset. JAFFE is a relatively small dataset consisting of 213 images. We have developed CNN models with dropout layers interspersed between convolution layers to avoid overfitting of the model. For this model, we have used the RMSProp optimizer and cross-entropy loss function. Model is trained for 30 epochs with a batch size equal to 8. The validation accuracy achieved on the JAFFE dataset is 84.37%. The confusion matrix for the JAFFE dataset is shown in Fig. 10.

Fig. 10 CNN model 2 confusion matrix

The precision, recall and f1 score achieved for model 2 is 91%, 84% and 86% respectively, shown in Table 3.

Table 3: CNN Model 2 Classification Report

	precision	recall	f1-score	support
Anger	1.00	0.83	0.91	6
Disgust	0.83	1.00	0.91	5
Fear	1.00	0.80	0.89	10
Happy	0.60	1.00	0.75	3
Neutral	1.00	0.75	0.86	4
Sad	0.75	0.75	0.75	4
Surprise	0.00	0.00	0.00	0
Accuracy			0.84	32
Macro Avg	0.74	0.73	0.72	32
Weighted Avg	0.91	0.84	0.86	32

4-2-3- Experiments with CK+ Dataset

The CNN model 3 we have developed utilizes the CK+ dataset for experimentation. While working with other datasets, we observed that the perception and demonstration of emotion vary from person to person. So, to avoid misclassification errors, we have used the CK+ dataset to make our architecture enrich with a dataset. All the images from the dataset are resized to dimension (48,48,3). We have used an RMSProp optimizer with default learning rate and cross-entropy as loss function. We have achieved 73.09% test accuracy. The confusion matrix is shown in Fig. 11. It is notable from Table 4 that, training and validation sets are unbalanced, which resulted in poor classification for some classes. From Table 4 we can observe that 'Contempt' emotion is hard to classify. It is either classified as 'Anger' or 'Disgust'. 'Contempt' emotion is part of only CK+ dataset, so for experimentation purposes, we have ignored the 'Contempt' emotion. Happy emotion is having good prediction accuracy than other emotions.

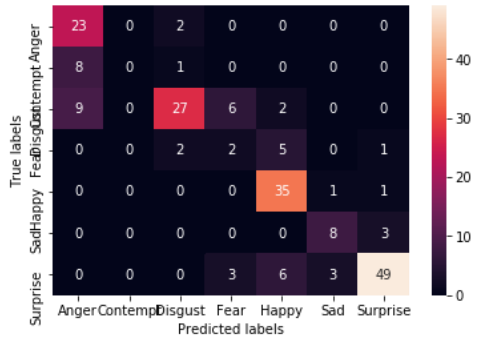

Fig. 11 CNN model 3 confusion matrix

Table 4: CNN Model 3 Classification Report

	precision	recall	f1-score	support
Anger	0.57	0.92	0.71	25
Contempt	0.00	0.00	0.00	9
Disgust	0.84	0.61	0.71	44
Fear	0.18	0.20	0.19	10
Happy	0.73	0.95	0.82	37
Sad	0.67	0.73	0.70	11
Surprise	0.91	0.80	0.85	61
Accuracy			0.73	197
Macro Avg	0.56	0.6	0.57	197
Weighted Avg	0.73	0.73	0.72	197

To validate the visual cues architecture, we have created our dataset named 'Emo-CSE Database'. The dataset

consists of the facial expressions of 43 plus student participants from the Computer Science and Engineering department. For the experiments, participants were asked to click their picture in a natural environment and annotate it with respective emotions. Most of the pictures were taken using a smartphone camera with no standard pre-settings. The samples from the dataset are shown in Fig. 12.

Fig. 12 Emo-CSE database

Overall accuracy achieved on the validation dataset is 76.24 %. Table 5. shows the performances of models for visual cues.

Table 5: Visual Cue Model Performances

Method	Accuracy (%)
CNN Model 1	71.25
CNN Model 2	84.37
CNN Model 3	73.09
Weighted Fusion	77.24

4-3- Experiments on Audio Cues

4-3-1- Experiments with RAVDESS Dataset

RAVDESS dataset contains audio-video files of song and speech data. We have considered only speech data for the experimentation. From the input speech data, we extracted 40 MFCC features of audio and used them to train the LSTM network. We trained the network for 100 epochs with Adam optimizer and cross-entropy as loss function. LSTM model achieves 79.51% test accuracy. The confusion matrix is shown in Fig. 13. Table 6 depicts the classification report for RAVDESS test set.

4-3-2- Experiments with TESS Dataset

For the TESS dataset, we have used the CNN model to process the audio files. 40 MFCC features are extracted from the input audio file with shape (40,1) and fed into a 2D convolution layer. We have a sparse categorical cross entry loss function and Adam optimizer. The model is trained for 50 epochs to achieve 78.37% test accuracy. The confusion matrix and classification report are shown Fig. 14 and Table 7. The summary of model performances on audio cues is shown in Table 8.

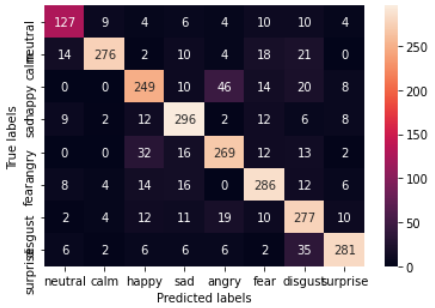

Fig. 13 LSTM model confusion matrix

Table 6: LSTM Model Classification Report

	precision	recall	f1-score	support
Neutral	0.77	0.73	0.75	174
Calm	0.93	0.80	0.86	345
Happy	0.75	0.72	0.73	347
Sad	0.80	0.85	0.82	347
Angry	0.77	0.78	0.78	344
Fear	0.79	0.83	0.81	346
Disgust	0.70	0.80	0.75	345
Surprise	0.80	0.82	0.85	344
Accuracy			0.8	2592
Macro Avg	0.80	0.79	0.79	2592
Weighted Avg	0.80	0.80	0.80	2592

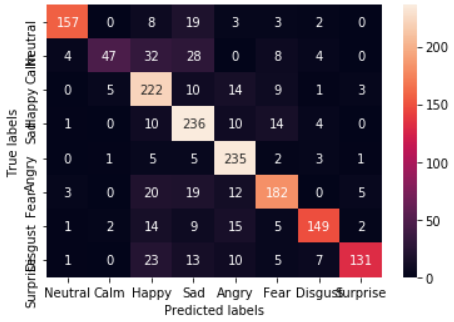

Fig. 14 CNN model for TESS confusion matrix

Table 7: CNN Model Classification Report

	precision	recall	f1-score	support
Neutral	0.94	0.84	0.87	192
Calm	0.85	0.38	0.53	123
Happy	0.66	0.84	0.74	264
Sad	0.70	0.86	0.77	275
Angry	0.79	0.93	0.85	252

Fear	0.80	0.76	0.78	241
Disgust	0.88	0.76	0.81	197
Surprise	0.92	0.69	0.79	190
Accuracy			0.78	1734
Macro Avg	0.82	0.75	0.77	1734
Weighted Avg	0.80	0.78	0.78	1734

Table 8: Audio Cue Model Performances

Method	Accuracy (%)
LSTM	79.51
CNN	78.37

4-4- Experiments on Text Cues

For text emotion recognition we have used a BERT transformer with the datasets mentioned in the above section. BERT uses an attention mechanism to learn the contextual relationship between words in sentences. The network is trained with a learning rate of 0.00002 and batch size of 6 for 2 epochs. We used the Ktrain library to implement a BERT transformer. For textual cue emotion recognition, we achieved 82.26% test accuracy. The confusion matrix and classification report are displayed in Fig. 14 and Table 9.

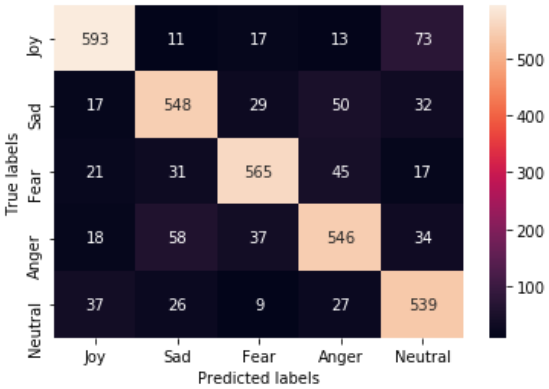

Fig. 14 Textual cue confusion matrix

Table 9: Textual cue Classification Report

	precision	recall	f1-score	support
Joy	0.86	0.84	0.85	707
Sad	0.81	0.81	0.81	676
Fear	0.86	0.83	0.85	679
Anger	0.80	0.79	0.79	693
Neutral	0.78	0.84	0.81	638
Accuracy			0.82	3393
Macro Avg	0.82	0.82	0.82	3393
Weighted Avg	0.82	0.82	0.82	3393

The proposed framework provides average accuracy of 79.48% for emotional distress detection from personalized affective cues when tested on e'NTERFACE dataset. The accuracy metrics of all the modalities are displayed in Table 10.

Table 10: Audio Cue Model Performances

Method	Accuracy (%)
Video cues	77.24
Audio cues	78.94
Text cues	82.36
Hierarchical Fusion	79.48

5- Conclusions

In this paper, we propose a hierarchical weighted framework for emotional distress detection using personalized affective cues i.e., facial expressions, audio signals and textual data. For the face affective state detection, we have utilized a multiple model approach with different datasets. We have also taken the advantage of transfer learning with ResNet50 to extract high-level features from video frames. Facial affective state recognition is challenging, considering variability in dataset and emotion demonstration by an individual. The second cue is an audio cue. We extracted low-level acoustic features of audio and stored them in a matrix to process it using two different approaches. We employed both CNN and LSTM capabilities separately for audio cues, which proved to be better than the combined CNN+LSTM approach. Text cue is last but not the least. For textual cues, we take advantage of the transferred learning using the BERT transformer to extract detailed word embeddings. All the cues are fused to generate a final affective state of the person. Finally, emotional distress is detected by analyzing the observed emotions. Experiments on multiple challenging datasets validate that our method is efficient and viable. For future work, the proposed algorithms can be deployed on smartphones for the self-assessment of emotional distress before consulting the clinical practitioners. The proposed algorithm can be combined with a clinical distress assessment questionnaire for effective results. Combining the results of our proposed architecture with clinical evidence will help in diagnosing mental disorders in the early stage.

Acknowledgements

I would like to acknowledge my supervisor Dr. Rekha Sugandhi, MIT ADT University, Pune, India for her

valuable support, insights and significant contribution in this research work.

References

[1] Gu Simeng, Wang Fushun, Patel Nitesh P., Bourgeois James A., Huang Jason H, "A Model for Basic Emotions Using Observations of Behavior in Drosophila," Frontiers in Psychology, vol. 10, 2019, pp.781.

[2] Rana R, Latif S, Gururajan R, Gray A, Mackenzie G, Humphris G, Dunn J, "Automated screening for distress: A perspective for the future," The European Journal of Cancer Care, vol. 28(4), 2019, pp. 1-13.

[3] Riba, M. B. et al., "Distress Management Version 3.2019 ," NCCN Clinical Practice Guidelines in Oncology, Journal of the National Comprehensive Cancer Network, vol.17(10), 2019, pp.1229–1249.

[4] A. Mehrabian and S.R. Ferris, "Inference of attitudes from nonverbal communication in two channels," International Journal of consulting psychology, vol. 31(3), 1967, pp. 248–252.

[5] T. Thomas, M. Domínguez and R. Ptucha, "Deep independent audio-visual affect analysis," in 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), 2017, pp. 1417-1421.

[6] J. A. Miranda, M. F. Canabal, J. M. Lanza-Gutiérrez, M. P. García and C. López-Ongil, "Toward Fear Detection using Affect Recognition," in 2019 XXXIV Conference on Design of Circuits and Integrated Systems (DCIS), November 2019, pp. 1-4.

[7] Lang He, Dongmei Jiang, and Hichem Sahli. "Multimodal depression recognition with dynamic visual and audio cues," Proc. 2015 International Conference on Affective Computing and Intelligent Interaction, IEEE Computer Society, USA, 2015, pp. 260–266.

[8] Guangxia Xu, Weifeng Li, Jun Liu, "A social emotion classification approach using multi-model fusion," Future Generation Computer Systems, vol. 102, 2020, pp. 347-356.

[9] Zeinab Farhoudi, Saeed Setayeshi, "Fusion of deep learning features with mixture of brain emotional learning for audio-visual emotion recognition," Speech Communication," vol. 127, 2021, pp. 92-103.

[10] Do LN., Yang HJ., Nguyen, HD. et al. "Deep neural network-based fusion model for emotion recognition using visual data," Journal of Supercomputing, vol.77, 2021, pp. 1-18.

[11] Neha Jain, Shishir Kumar, Amit Kumar, Pourya Shamsolmoali, Masoumeh Zareapoor, "Hybrid deep neural networks for face emotion recognition," Pattern Recognition Letters, vol. 115, 2018, pp. 101-106.

[12] Heysem Kaya, Furkan Grpnar, and Albert Ali Salah, "Video-based emotion recognition in the wild using deep transfer learning and score fusion," Image Vision Computing, vol. 65, 2017, pp. 66–75.

[13] Shiqing Zhang, Xin Tao, Yuelong Chuang, Xiaoming Zhao, "Learning deep multimodal affective features for spontaneous speech emotion recognition, Speech Communication," vol. 127, 2021, pp. 73-81.

[14] Jingwei Yan, Wenming Zheng, Zhen Cui, Chuangao Tang, Tong Zhang, Yuan Zong, "Multi-cue fusion for emotion recognition in the wild," Neurocomputing, vol. 309, 2018, pp. 27-35.

[15] Ilyes Bendjoudi, Frederic Vanderhaegen, Denis Hamad, Fadi Dornaika, "Multi-label, multi-task CNN approach for context-based emotion recognition," Information Fusion, November 2020, in press.

[16] Man Hao, Wei-Hua Cao, Zhen-Tao Liu, Min Wu, Peng Xiao, "Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features," Neurocomputing, vol. 391, 2020, pp. 42-51.

[17] Tzirakis, P., Trigeorgis, G., Nicolaou, M.A., Schuller, B., Zafeiriou, S., "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks," IEEE Journal of Selected Topics in Signal Processing, vol. 11, 2017, pp. 1301-1309.

[18] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling, Knowledge-Based Systems," vol. 161, 2018, pp. 124-133.

[19] Soujanya Poria, Erik Cambria, Amir Hussain, Guang-Bin Huang, "Towards an intelligent framework for multimodal affective data analysis," Neural Networks, vol. 63, 2015, pp. 104-116.

[20] Yaxiong Ma, Yixue Hao, Min Chen, Jincai Chen, Ping Lu, Andrej Košir, "Audio-visual emotion fusion (AVEF): A deep efficient weighted approach," Information Fusion, vol. 46, 2019, pp. 184-192.

[21] Jie Guo, Bin Song, Peng Zhang, Mengdi Ma, Wenwen Luo, Junmei lv, "Affective video content analysis based on multimodal data fusion in heterogeneous networks," Information Fusion, vol. 51, 2019, pp. 224-232.

[22] Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, Amir Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," Neurocomputing, vol. 174, Part A, 2016, pp. 50-59.

[23] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera and G. Anbarjafari, "Audio-Visual Emotion Recognition in Video Clips," IEEE Transactions on Affective Computing, vol. 10, 2019, pp. 60-75.

[24] S. Zhang, S. Zhang, T. Huang, W. Gao and Q. Tian, "Learning Affective Features with a Hybrid Deep Model for Audio–Visual Emotion Recognition," IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, 2018, pp. 3030-3043.

[25] Li, R., Liu, Z., "Stress detection using deep neural networks.," BMC Medical Informatics and Decision Making vol. 20, 2020, pp. 285.

[26] P. Bobade and M. Vani, "Stress Detection with Machine Learning and Deep Learning using Multimodal Physiological Data," in 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 51-57.

[27] Zhang, H., Feng, L., Li, N., Jin, Z., & Cao, L., "Video-Based Stress Detection through Deep Learning.," Sensors, vol. 20(19), 2020, pp. 5552.

[28] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," Neural Networks, Special Issue on Deep Learning of Representations, vol. 64, 2015, pp. 59-63.

[29] Lyons, Michael, Kamachi, Miyuki, & Gyoba, Jiro, "The Japanese Female Facial Expression (JAFFE) Dataset," Zenodo. 1998.

[30] Lucey et al., "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified

expression," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 2010, pp. 94-101.

[31] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016.

[32] Mehmet Berkehan Akçay, Kaya Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers, Speech Communication," vol. 116, 2020, pp. 56-76.

[33] Zheng, F., Zhang, G. & Song, Z., "Comparison of different implementations of MFCC," Journal of Computer Science & Technology, vol.16, 2001, pp. 582–589.

[34] Livingstone SR, Russo FA, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," PLoS ONE vol 13(5), 2018.

[35] Pichora-Fuller, M. Kathleen; Dupuis, Kate, "Toronto emotional speech set (TESS)", Scholars Portal Dataverse, V1, 2020.

[36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova; "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, June 2019.

[37] I. Mureşan, A. Stan, M. Giurgiu and R. Potolea, "Evaluation of sentiment polarity prediction using a dimensional and a categorical approach," in 7th Conference on Speech Technology and Human - Computer Dialogue (SpeD), 2013, pp. 1-6.

[38] Diman Ghazi, Diana Inkpen & Stan Szpakowicz, "Detecting Emotion Stimuli in Emotion-Bearing Sentences". in 16th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2015), Cairo, Egypt.

[39] Li Yanran, Su Hui, Shen Xiaoyu, Li Wenjie, Cao Ziqiang, Niu Shuzi; "DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset," in Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers),2017,nov,Asian Federation of Natural Language Processing,Taipei, Taiwan,Pages pp. 986-995.

[40] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, G. Rigoll, "LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework," Image and Vision Computing, vol. 31, 2013, pp. 153-163.

[41] S. Chen and Q. Jin. "Multi-Modal Dimensional Emotion Recognition Using Recurrent Neural Networks", Proc. 5th International Workshop on Audio/Visual Emotion Challenge. AVEC '15. Brisbane, Australia: Association for Computing Machinery, 2015, pp. 49-56.