Attention-based Multi-modal Sentiment Analysis and Emotion Detection in Conversation using RNN

Mahesh G. Huddar[1,3]*, Sanjeev S. Sannakki[2,3], Vijay S. Rajpurohit[2,3]

[1] Department of Computer Science and Engineering, Hirasugar Institute of Technology, Nidasoshi, Belagavi (India)
[2] Department of Computer Science and Engineering, Gogte Institute of Technology, Belagavi (India)
[3] Visvesvaraya Technological University, Belagavi (India)

Received 6 June 2019 | Accepted 26 October 2020 | Published 1 December 2020

Abstract

The availability of an enormous quantity of multimodal data and its widespread applications, automatic sentiment analysis and emotion classification in the conversation has become an interesting research topic among the research community. The interlocutor state, context state between the neighboring utterances and multimodal fusion play an important role in multimodal sentiment analysis and emotion detection in conversation. In this article, the recurrent neural network (RNN) based method is developed to capture the interlocutor state and contextual state between the utterances. The pair-wise attention mechanism is used to understand the relationship between the modalities and their importance before fusion. First, two-two combinations of modalities are fused at a time and finally, all the modalities are fused to form the trimodal representation feature vector. The experiments are conducted on three standard datasets such as IEMOCAP, CMU-MOSEI, and CMU-MOSI. The proposed model is evaluated using two metrics such as accuracy and F1-Score and the results demonstrate that the proposed model performs better than the standard baselines.

Keywords

Attention Model, Interlocutor State, Contextual Information, Emotion Detection, Multimodal Fusion, Sentiment Analysis.

DOI: 10.9781/ijimai.2020.07.004

I. Introduction

THE main aim of automatic sentiment and emotion analysis in conversational videos is to analyze and detect sentiment and emotional state of a participant in conversational videos. Due to the recent advancements in Internet technologies and social media networks, the users post their reviews, about a service or a product in the form of conversational videos on social media platforms, such as Twitter, Flicker, YouTube, and Facebook, etc. Recently, multimodal sentiment and emotion analysis from the conversation has become an interesting research topic due to its widespread applications in areas such as health-care assistant devices, education, dialogue understanding, human-computer interaction, and human resource management. In prior work, the unimodal features from the available modalities were extracted, and then the unimodal features are fused to form the multimodal feature vector. For multimodal fusion, there are three options, early fusion (feature concatenation), model-based fusion, or late (decision) fusion. In feature concatenation, the features from individual modalities are concatenated to get the multimodal feature vector.

Recently, many approaches were proposed for utterance level sentiment and emotion analysis [1], [2]. In late fusion, the feature vectors from individual modalities are modeled using the classical classifiers. The output of the classifiers on the unimodal feature vector is fused using an ensemble approach [3]. These fusion strategies perform fairly well but cannot accommodate the contextual information among the utterances and interlocutor state of the participant. More recently attention based contextual fusion and contextual cross modality fusion strategies show promising results. In the contextual fusion technique, the bidirectional recurrent neural network (RNN) was used to extract the context between the utterances of a video [4]. In contextual cross-modality fusion along with contextual information, the importance of modality is considered in multimodal fusion [5]. In [6] dynamic fusion is performed by paying attention at each time step. Evolutionary computing-based multi-layer feature optimization is used to improve the overall accuracy of classification in [7].

The sentiment or emotional state of the particular participant in the conversation is not considered for analysis in these models. Hence the existing models fail to capture the contextual information among the utterances and flow of conversation. But in reality, the contextual state and sentiment or emotion of a particular party does add a lot of value to the overall result. The proposed model believes that the sentiment or emotional state of an utterance mainly depends on the interlocutor state of the participant, the previous emotional state of the participant, and context between the utterances [8]. By incorporating the interlocutor state of the particular participant and context between the utterances, the results of the proposed method outperform the baselines by over 2%.

The main contributions of the proposed model are,

- An effective multimodal sentiment and emotion analysis technique is proposed to extract the contextual information among the utterances and accommodate the interlocutor state of a particular participant in the conversation.

* Corresponding author.
E-mail address: mailtomgh1@gmail.com

- The pair-wise attention-based mechanism is used to understand the relationship and importance of modalities before fusion.
- The proposed model effectively captures the sentiment or emotional state of the participant in the conversation.
- The model is tested and validated on three standard datasets and the results are compared against the standard baselines for multimodal sentiment and emotion analysis in conversational videos.

The structure of the remaining sections of the article is as follows: the important work carried-out in multimodal sentiment and emotion analysis, context extraction between the utterances and traditional techniques in multimodal fusion are described in Section II. The proposed attention-based multimodal sentiment and emotion analysis in the conversation using the RNN model is presented in Section III. The experimental setup, results on three standard datasets, and comparison of results against a standard baseline of the proposed model are presented in Section IV. Finally, future work in multimodal affective computing in conversational videos is presented and concludes the paper in Section V.

II. Related Work

Sentiment analysis and Emotion detection in conversation are popular research topics in multimodal affective computing [9] because of their applications in various areas such as sentiment analysis, health-care assistance devices, recommendation systems, education, human-computer interaction, etc. [10]. The multimodal data has information in three modes such as text (transcribed audio), audio, and video. The traditional multimodal sentiment analysis and emotion detection technique extracts the unimodal features from the three modalities, use either feature level (early) fusion [11] [12] or decision level (late) fusion [13] [14] [15] or hybrid fusion [16] to merge effective information from different modalities.

An utterance is a segment or a part of the video (may not be a complete sentence) and video reviews contain a sequence of such multiple utterances. In utterance or segment level sentiment and emotion classification, each segment of a video is analyzed and assigned a label [17]. Recently, many approaches were proposed for analyzing sentiment and detecting emotion at the utterance level [1], [2]. In [18] authors extracted acoustic, lexicon, and visual features and used an ensemble approach to ensemble classification of SVM classifier. Their proposed ensemble approach achieves better results than conventional methods. Authors in [19] fused acoustic and linguistic cues at feature level using 3-D activation valance for emotion recognition. In [20] authors extracted textual, speech, and visual features using convolutional neural networks. They analyzed sentiment and emotion using multiple kernel learning.

In [21] acoustic information and visual cues are fused to model multimodal emotion recognition system and contextual information is used for sentiment and emotion analysis. In recent works on multimodal sentiment and emotion analysis in conversational videos, each utterance of a video is processed sequentially using RNN. The model proposed in [8] propagates the context among the utterances and sequential information to the next utterance. They use bidirectional recurrent neural networks [22] to extract the context between the utterances and feed the information sequentially. DialogueRNN [23] uses an attention-based pooling approach to capture the context of a particular utterance in the conversation. However, this pooling based attention mechanism fails to consider participant information of particular utterance and its effect on other utterances. They use a global state and participant state for modeling multimodal emotion detection in conversation.

Other notable works include [24] [25] [26] where multimodal sentiment and emotion detection is addressed using deep learning-based models. Ghosal et al. [27] proposed a pair-wise attention-based method to understand the importance of individual modalities and the relationship between the modalities before fusion. The two-dimensional graph-based feature extraction methods using fuzzy logic are discussed in [28] [29] and [30]. The PRAAT[1] software was used to extract the emotional state from voice [31]. The proposed model considers context between the utterances, the interlocutor state of a participant, and previous emotion state to effectively model the multimodal sentiment and emotion analysis system in conversational videos.

III. Proposed Methodology

The proposed attention-based multimodal sentiment and emotion analysis in the conversation using RNN is discussed in detail in this section. The overview of the proposed model is:

- First, the utterance level features of individual modalities such as acoustic, textual, and visual features are extracted.
- The pair-wise attention-based mechanism is used to understand the relationship and importance of modalities before fusion.
- The gated recurrent unit (GRU), a variant of RNN, is used to model the interlocutor state of the participant, context extraction, and emotion decoding.
- Bimodal and trimodal fusion are performed by considering the previous emotional state, the importance of individual modality, and interlocutor state. A trimodal representation of feature vector acts as an input for final sentiment or emotion prediction.

A. Dataset Used

The model is evaluated on three standard datasets, IEMOCAP [32], CMU-MOSEI [24] for multimodal emotion detection, and CMU-MOSI [33] for multimodal sentiment analysis.

1. IEMOCAP

IEMOCAP dataset is a collection of 12-hours of two-way acted dyadic conversations among multiple speakers. The conversational video is divided into multiple opinion segments called utterances. Each of the utterances is annotated with emotion labels such as anger, sadness, excitement, happiness, fear, neutral, and surprise. Videos with angry, happy, sad, excited, frustrated, and neutral are considered to compare against the state of the art models.

2. CMU-MOSEI

The CMU-MOSEI dataset contains 3228 videos with 23453 small segments called utterances from 1000 speakers collected from YouTube. CMU-MOSEI is a transcribed, gender-balanced, properly punctuated dataset. The average number of segments per video is 7.3 and the average length of each segment is 7.28 seconds. The total number of words and unique words in utterances are 447143 and 23026 respectively. The dataset is manually labeled with 6 emotions such as anger, disgust, fear, happiness, sadness, and surprise.

3. CMU-MOSI

There are 93 videos with 2199 utterances in CMU-MOSI dataset where 89 speakers review various products and topics in English. The average length of a segment is 4.2 seconds and about 12 words per utterance. Each utterance is manually labeled by 5 assessors with a score ranging from -3 and +3. The average of these 5 assessors is taken as sentiment polarity. The Video/Utterance level Train-Test

[1] https://www.fon.hum.uva.nl/praat/

distributions of CMU-MOSEI, CMU-MOSI, and IEMOCAP datasets are shown in Table I. The Label distribution statistics of CMU-MOSI and IEMOCAP datasets are given in Table II and Table III respectively.

TABLE I. Video/Utterance Level Train-Test Distribution of CMU-MOSI, CMU-MOSEI, and IEMOCAP Dataset

	Videos		Utterances	
	Train	Test	Train	Test
CMU-MOSI	62	31	1447	752
CMU-MOSEI	2583	646	18051	4625
IEMOCAP	120	31	5810	1623

TABLE II. Label Distribution Statistics of CMU-MOSI

	Positive	Negative
CMU-MOSI	1176	1023

TABLE III. Label Distribution Statistics of IEMOCAP

	Neutral	Happy	Sadness	Anger	Frustrated	Excited
IEMOCAP	1708	648	1084	1103	1849	1041

B. Feature Extraction

This section discusses the steps followed in extracting features from acoustic, text, and visual modalities.

1. Audio Feature Extraction

OpenSMILE [34] open-source tool is used for acoustic feature extraction from CMU-MOSI, CMU-MOSEI, and IEMOCAP datasets. The acoustic features are extracted at a frame rate of 30Hz and 100ms sliding window. The dimension of utterance level features for acoustic modality is 73, 73, and 384 for CMU-MOSI, IEMOCAP, and CMU-MOSEI datasets respectively.

Let f_{ai} be the feature vector of i^{th} segment then, the acoustic feature vector f_a is represented by,

$$f_a = < f_{a1}, f_{a2}, f_{a3}, \dots \dots, f_{an} > \quad (1)$$

Where n is the number of segments or utterances.

2. Textual Feature Extraction

Features from text (transcribed text) modality are extracted from each utterance using Convolutional Neural Networks (CNN) [35] from CMU-MOSI and IEMOCAP datasets. First, each utterance is represented as word2vec vectors [36] to understand the context in the text. These Word2Vec vectors are processed using 3 convolutional layers. The three layers have feature maps of size 50, 75, and 100 with filters of sizes 2, 3, and 2 respectively. Max-pooling of a 2x2 window size is used after every convolutional layer. The fully connected layer receives input from the convolution layer and output is fed to a softmax classifier. The fully connected layer has 600 neurons with ReLU [37] activation function. The softmax output of the convolutional neural network (CNN) is used as the textual features. GloVe embedding's used for extracting textual features from the CMU-MOSEI dataset. The dimension of utterance level features for textual modality is 100 for CMU-MOSI, IEMOCAP datasets, and 300 for CMU-MOSEI dataset.

Let f_{ti} be the feature vector of i^{th} segment then, the textual feature vector f_t is represented by,

$$f_t = < f_{t1}, f_{t2}, f_{t3}, \dots \dots, f_{tn} > \quad (2)$$

Where n is the number of segments or utterances.

3. Visual Feature Extraction

In the past 3D convolutional neural networks have been successfully used for object detection and classification [38]. The results presented in [38], outperform the traditional object tracking and detection, and motivate us to adopt 3D-CNN in our work. Visual features are extracted using 3D-CNN from CMU-MOSI and IEMOCAP datasets and Facet[2] tool from the CMU-MOSEI dataset. The dimension of utterance level features for visual modality is 100 for CMU-MOSI, IEMOCAP datasets, and 35 for CMU-MOSEI dataset.

Let f_{vi} be the feature vector of i^{th} segment then, the visual feature vector f_v is represented by,

$$f_v = < f_{v1}, f_{v2}, f_{v3}, \dots \dots, f_{vn} > \quad (3)$$

Where n is the number of segments or utterances.

C. Problem Statement

Let P_1 and P_2 be the two participants in the conversation. The $u_1, u_2 \dots u_n$ are the utterances uttered by either of the participants P_1 and P_2 with sentiment score and one of the emotion labels such as happy, sad, anger, surprise, disgust, and fear is assigned to the utterances. As each of the utterances is uttered by either of the participants in the conversation, this allows capturing the average sentiment of the participant in sentiment score or emotion label calculation. Also, it avoids misclassification due to long pauses by the participant in the conversation. Let u_t be the t^{th} utterance uttered by the party P_1 or P_2 at timestamp t, which is represented by three modalities such as text, visual and acoustic,

$$u(p)_t = < t_t, v_t, a_t > \quad (4)$$

where t_t, v_t, and a_t are textual, visual and acoustic feature vectors of the t^{th} utterance at timestamp t and p Є P1, P2.

The objective function of the problem is to accept the feature vector from three modalities of an utterance, cumulative context representation of the conversation and emotional state of the previous participant, and output the sentiment score and associated emotion label.

D. Proposed Model Description

The sentiment or emotion of an utterance depends on the cumulative contextual state of the conversation, the interlocutor state, and the sentiment or emotional state of the previous participant. Hence the proposed model considers the cumulative context and emotion of participants to predict the sentiment or emotional state of an utterance. The proposed model has three branches of recurrent neural networks (RNN) to capture the participant interlocutor state, cumulative context, and sentiment or emotional state of the participant. Each modality uses one RNN to capture participant dyadic information and another set of RNN's are used to capture the sentiment or emotional state of the participant. One RNN is used to capture the cumulative contextual information. A weighted-pooling based pairwise attention-based mechanism is performed to understand the relative importance of individual modalities before fusion. Finally, two-two modalities and then all modalities are fused to form a trimodal representation of feature vector for predicting the sentiment score or emotion label of an utterance.

1. Interlocutor State

The interlocutor state of the network captures and keeps track of the state of the participant involved in the multimodal conversation. The network has nxm number of RNN's, where n is the number of participants and m is the number of modalities. The output of the interlocutor state is the input for updating cumulative contextual vector and emotion or sentiment prediction of the utterance. Initially, the interlocutor state is initialized to the null vector. For the utterance at timestamp t the interlocutor state i_t of a particular modality is updated i_{t+1} using feature

[2] https://goo.gl/1rh1JN

representation of particular modality at timestamp t (that is $f(t)_t$ or $f(a)_t$ or $f(v)_t$) and attentive cumulative contextual vector representation until timestamp t (that is $C(t)_t$ or $C(a)_t$ or $C(v)_t$). The purpose of using the cumulative contextual vector along with utterance representation is to understand the contextual information of conversation until that timestamp. The steps in the interlocutor state update are described using the following formula and shown in Fig. 1.

$$i(m)_t = \text{Interlocutor}((f(m)_t \oplus C(m)_t), i(m)_{t-1}) \quad (5)$$

where $\oplus$ represents concatenation operator and m is the modality with values either t or a or v.

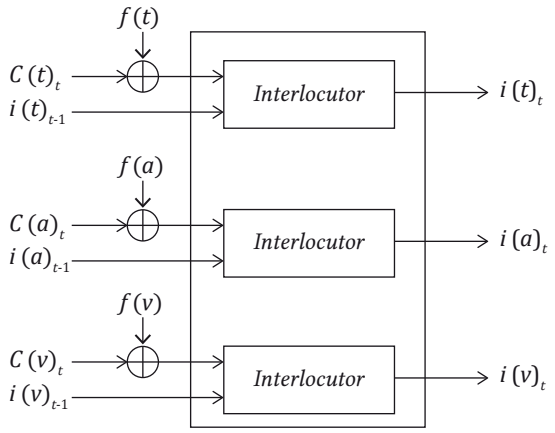

Fig. 1. Interlocutor State Update at timestamp t.

2. Cumulative Context

In conversational sentiment analysis and emotion detection, to determine the sentiment or emotional state of an utterance at timestamp t, the preceding utterances at time < t can be considered as its cumulative context. The interlocutor state of the previous utterance (that is $i(t)_{t-1}$ or $i(a)_{t-1}$ or $i(v)_{t-1}$) and utterance level modality representation at timestamp t (that is $f(t)_t$ or $f(a)_t$ or $f(v)_t$) are used to change the cumulative context vector representation from c_{t-1} to c_t. This helps to understand the dependencies between the utterances and participants. The steps in the cumulative context state update are described using the following formula and shown in Fig. 2.

$$c(m)_t = \text{Context}((f(m)_t \oplus i(m)_{t-1}), c(m)_{t-1}) \quad (6)$$

where $\oplus$ represents concatenation operator and m is the modality with values either t or a or v.

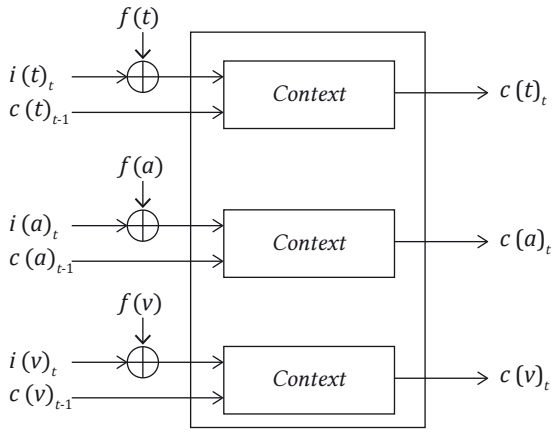

Fig. 2. Cumulative Context state update at timestamp t.

Weighted pooling based attention is performed over cumulative context vector representation until timestamp t.

$$C(m)_t = \frac{e^{c(m)_t}}{\sum_{k=1}^{t} e^{c(m)_k}} \quad (7)$$

Where, $C(m)_t$ is the attentive cumulative contextual vector.

3. Emotion State

The emotional state network is used to decode the sentiment or emotional information encoded by interlocutor state RRN. The previous emotion state output (that is $e(t)_{t-1}$ or $e(a)_{t-1}$ or $e(v)_{t-1}$) and interlocutor state sentiment or emotional information (that is $i(t)_t$ or $i(a)_t$ or $i(v)_t$) are the input to emotion state RNN at timestamp t. Weighted pooling based pair-wise attention is performed on the output produced by emotion state RNN to produce the relevant sentiment or emotion label. The steps in the emotion state update are described using the following formula and shown in Fig. 3.

$$e(m)_t = \text{emotion}(i(m)_t, e(m)_{t-1}) \quad (8)$$

where $\oplus$ represents concatenation operator and m is the modality with values either t or a or v.

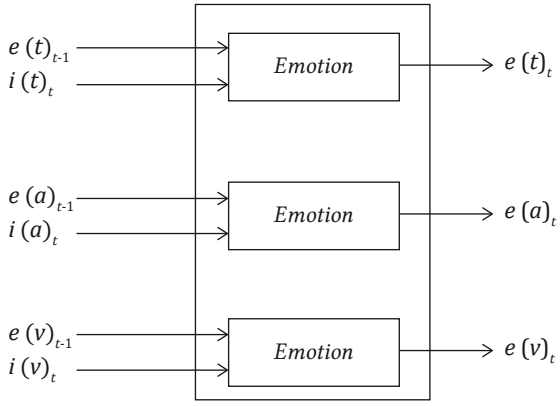

Fig. 3 Emotion State update at timestamp t.

4. Weighted Pooling Based Pair-Wise Attention and Bimodal Fusion

For each timestamp t, the emotion state network produces emotion vectors for each modality such as $e(t)_t$, $e(a)_t$ and $e(v)_t$. Weighted pooling based pair-wise attention [4] is performed between two-two emotion vectors at a time to get bimodal representation emotion vectors. Let X and Y be the two emotion state outputs produced by the emotion state network at timestamp t, then the weighted pooling based pair-wise attention mechanism is performed as follows:

$$M(t)_1 = X.Y^T \quad \text{and} \quad M(t)_2 = Y.X^T \quad (9)$$

$$W(t)_1 = \frac{e^{M(t)_1}}{\sum_{k=1}^{t} e^{M(k)_1}} \quad (10)$$

$$W(t)_2 = \frac{e^{M(t)_2}}{\sum_{k=1}^{t} e^{M(k)_2}} \quad (11)$$

$$O(t)_1 = W(t)_1 \,.\, Y \quad \text{and} \quad O(t)_2 = W(t)_2 \,.\, X \quad (12)$$

$$A(t)_1 = O(t)_1 \odot X \quad \text{and} \quad A(t)_2 = O(t)_2 \odot Y \quad (13)$$

$$B\_Fusion(XY)_t = A(t)_1 \oplus A(t)_2 \quad (14)$$

Where B_Fusion is the bimodal fusion at timestamp t.

The pair-wise matching matrices at timestamp t are calculated in equation (9), then the probability distribution scores (weights)

of each modality are calculated in equation (10) and (11). Modality specific attentive representations are calculated in equation (12). An important component among the multiple modalities and utterances is calculated by performing element-wise matrix multiplication as shown in equation (13). Attentive matrix representations are then concatenated to produce bimodal representation at timestamp t as shown in equation (14). The steps in Weighted Pooling based pair-wise attention and bimodal fusion at timestamp t are shown in Fig. 4.

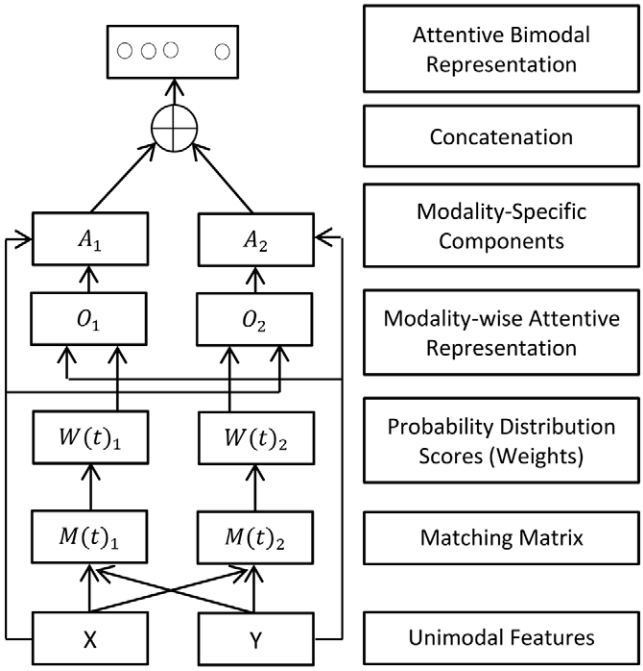

Fig. 4. Weighted Pooling based Pair-Wise Attention and Bimodal Fusion at timestamp t where X, Y Є {T, A, V}.

5. Trimodal Fusion

The bimodal attentive representation and emotional state of the utterance are used to get the trimodal representation. The bimodal attentive representation and output of emotion state RNN at timestamp t is concatenated to form the final trimodal attentive representation at timestamp t. The trimodal fusion at timestamp t is shown in Fig. 5.

$$e(tav)_t = e(t)_t \oplus e(a)_t \oplus e(v)_t \oplus B\_Fusion(TA) \oplus B\_Fusion(TV) \oplus B\_Fusion(VA) \quad (15)$$

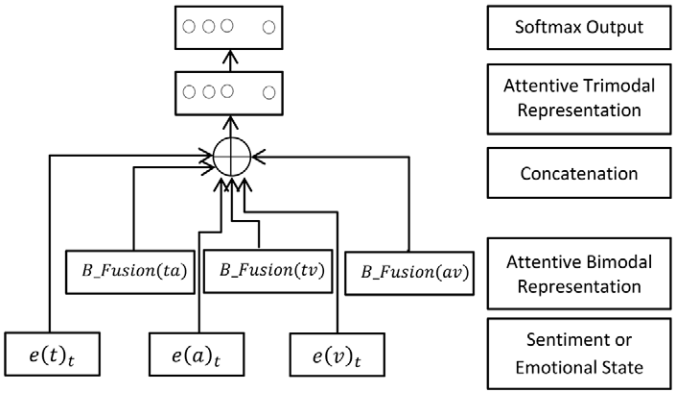

Fig. 5. Trimodal Fusion at timestamp t.

E. Classification and Training

The trimodal sentiment or emotional representation is fed to the softmax classifier to predict the testing label $\hat{y}$ for an utterance in the conversation. The softmax classifier takes the concatenated sentiment or emotion vector $e(tav)_t$ at timestamp t as an input. The softmax output is represented as,

$$p(y|U) = \text{softmax}\left(w^{(s)}(e(tav)_t) + b^{(s)}\right) \quad (16)$$

Where $w^{(s)}$ is the weight matrix, $b^{(s)}$ is the bias matrix, p is a predicted sentiment or emotion class.

$$\hat{y} = \underset{y}{\text{argmax}}\ p(y|U) \quad (17)$$

Where $\hat{y}$, is the predicted label of testing utterance.

The cross-entropy loss function L(θ) is used to train the model and is represented as,

$$L(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_i^j \log\hat{y}_i^j + \lambda\sum_{k=0}^{N}\theta_k^2 \quad (18)$$

where N is the number of utterances in training data. y_s and $\hat{y}_s$ are the true and predicted label of the s^{th} utterance. M is the number of categories (classes) and λ is the L2-regularization term. Adam [39] is used to optimize the cross-entropy loss function parameters due to its ability to adapt to the learning rate for each learning parameter. The proposed algorithm for attention-based multimodal sentiment and emotion analysis in the conversation using RNN is summarized in Table IV.

IV. Results Analysis and Discussion

The proposed attention-based multimodal sentiment and emotion analysis framework in the conversation using RNN is implemented in python using the PyTorch and tensor flow is used as backend. The model is evaluated on the Tesla K80 GPU with a 12GB RAM hardware configuration. The experiments are conducted on three standard datasets such as CMU-MOSI, CMU-MOSEI, and IEMOCAP. The experimental results of the proposed method are compared against the standard baselines such as [25] , [27], [40], [41], and [42]. The proposed model is evaluated using two metrics, classification accuracy, and F1-score. First, the results are obtained for the combination of two-two modalities such as text-audio, text-video, and audio-video, and then all three modalities with and without attention mechanism. The comparison of results of the proposed technique for sentiment analysis with and without attention is given in Tables V and VI. The results show that the attention-based model performs better than the standard baselines in all possible combinations of constituent modalities except for the audio-video combination on the CMU-MOSI dataset. The trimodal model performs better than the bimodal model. The emotion detection results on CMU-MOSEI and IEMOCAP datasets with and without attention are shown in Tables VII and VIII. The results show that the attention-based models are performing better than the standard baselines and model without attention except for the label happy in the CMU-MOSEI dataset. Fig. 6, Fig. 7, Fig. 8 and Fig. 9 show a comparison of the experimental results of the proposed method on CMU-MOSEI, IEMOCAP, CMU-MOSI datasets against standard baselines. On CMU-MOSI and CMU-MOSEI datasets the trimodal models are performing better than the bimodal and unimodal models, whereas A-V combination is performing the worst among all possible combination of models in sentiment classification. For emotion classification, the proposed model obtains the best results on the CMU-MOSEI dataset as it effectively uses all the available modalities and captures the contextual information since the availability of large dataset for training.

TABLE IV. ALGORITHM FOR PROPOSED MULTIMODAL SENTIMENT AND EMOTION CLASSIFICATION IN THE CONVERSATION USING RNN

1: *Procedure* $FeatureExtraction(U)$ 2: *for* i *in* 1 *to* N *do*: 3: $f(t)_i \longleftarrow audioFeatures(U_i)$ 4: $f(a)_i \longleftarrow textFeatures(U_i)$ 5: $f(v)_i \longleftarrow videoFeatures(U_i)$	*Procedure to extract unimodal features*
6: *Procedure* $InterlocutorState(t, m)$ 7: $i(m)_t = Intrerlocutor((f(m)_t \oplus C(m)_t), i(m)_{t-1})$ 8: *return* $(i(m)_t)$	*Procedure to update Interlocutor state at time t* $m \in \{t, a, v\}$
9: *Procedure* $ContextExtract(t, m)$ 10: $c(m)_t = Context((f(m)_t \oplus i(m)_{t-1}), c(m)_{t-1})$ 11: $\mathrm{C(m)_t} = \frac{e^{c(m)_t}}{\sum_{k=1}^{t} e^{c(m)_k}}$ 12: $return(C(m)_t)$	*Procedure to extract cumulative context* $m \in \{t, a, v\}$
13: *Procedure* $EmotionState(t, m)$ 14: $e(m)_t = Emotion(i(m)_t, e(m)_{t-1})$ 15: $return(e(m)_t)$	*Procedure to update Emotion state at time t* $m \in \{t, a, v\}$
16: *Procedure* $Attention(t, X, Y)$ 17: $M(t)_1 = X.\mathrm{Y^T}$ *and* $M(t)_2 = Y.\mathrm{X^T}$ 18: $W(t)_1 = \frac{e^{M(t)_1}}{\sum_{k=1}^{t} e^{M(k)_1}}$ 19: $W(t)_2 = \frac{e^{M(t)_2}}{\sum_{k=1}^{t} e^{M(k)_2}}$ 20: $O(t)_1 = W(t)_1 . Y$ *and* $O(t)_2 = W(t)_2 . X$ 21: $A(t)_1 = O(t)_1 \odot X$ *and* $A(t)_2 = O(t)_2 \odot Y$ 22: $return(A(t)_1 \oplus A(t)_2)$	*Procedure for weighted pooling based attention and fusion*
23: *Procedure* $B\_Fusion(t, X, Y)$ 24: $B(X, Y)_t = Attention(t, X, Y)$ 25: $retrun(B(X, Y)_t)$	*Procedure for Bimodal fusion at time t*
26: *Procedure* $T\_Fusion(t)$ 27: $e(tav)_t = e(t)_t \oplus e(a)_t \oplus e(v)_t \oplus B\_Fusion(TA) \oplus B\_Fusion(TV) \oplus B\_Fusion(VA)$ 28: $retrun\ (e(tav)_t)$	*Procedure for Trimodal fusion at time t*
29: *Procedure* $Classification\ (U, t)$ 30: *for* i *in* 1 *to* N *do*: 31: $p\ (y\|U) = softmax\ (w^{(s)}(e(tav)_t) + b^{(s)})$ 32: $\hat{y} = \underset{y}{argmax}\ p\ (y\|U)$ 33: $return\ (\hat{y})$	*Procedure for classification of utterance into discrete number of classes*
34: $FeatureExtraction(U)$	*Unimodal Feature Extraction*
35: *for* $X, Y \in \{t, a, v\}$ 36: $B(X, Y)_t \longleftarrow B\_Fusion(t, X, Y)$	*Bimodal Fusion* $X, Y \in \{t, a, v\}$
37: $e(tav)_t \longleftarrow T\_Fusion(t)$	*Trimodal Fusion*
38: $C_t \longleftarrow Classification\ (e(tav)_t)$	*Classification*

TABLE V. Experimental Results of the Proposed Method on CMU-MOSI Dataset Compared Against Standard Baselines, T for Text, A for Audio and V for Video

Modality	Poria et al. [40]	Zadeh et al. [41]	GRU - Without Attention		GRU - With Attention	
	Accuracy	Accuracy	Accuracy	F1-Score	Accuracy	F1-Score
T + A	73.7	71.1	76.79	77.62	79.71	73.38
T + V	74.1	73.7	79.35	78.54	80.14	73.40
A + V	68.4	67.4	65.51	67.77	66.28	67.79
T+A+V	74.1	73.6	80.02	73.73	80.62	74.33

TABLE VI. Experimental Results of the Proposed Method on CMU-MOSEI Dataset Compared Against Standard Baselines

Modality	Zadeh et al. [42]	Ghosal et al. [27]	GRU - Without Attention		GRU - With Attention	
	Accuracy	Accuracy	Accuracy	F1-Score	Accuracy	F1-Score
T + A	-	79.74	80.02	73.73	80.64	73.29
T + V	-	79.40	80.31	73.42	80.54	76.22
A + V	-	76.66	76.79	77.62	80.45	73.50
T+A+V	76.90	79.80	80.98	73.51	81.29	73.12

TABLE VII. Experimental Results of the Proposed Method on CMU-MOSEI Dataset Compared Against Standard Baselines with the T-A-V Combination of Modalities

Label	Ghosal et al. [27]		GRU - Without Attention		GRU - With Attention	
	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
Anger	62.60	72.80	81.49	74.20	83.58	76.10
Fear	62.00	89.90	88.92	84.83	91.20	87.01
Happy	66.30	66.30	58.51	44.30	59.79	44.74
Sad	60.40	66.90	76.93	67.86	78.90	69.60
Surprise	53.70	85.50	87.51	82.78	89.75	84.91
Disgust	69.10	76.60	88.92	84.83	91.20	87.01

TABLE VIII. Experimental Results of the Proposed Method on IEMOCAP Dataset Compared Against Standard Baselines with the T-A-V Combination of Modalities

Label	Ghosal et al. [27]		Hazarika et al. [25]		GRU - Without Attention		GRU - With Attention	
	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
Angry	64.7	65.2	68.2	68.2	73.1	66.6	75.2	68.5
Neutral	58.5	59.2	59.9	60.6	79.8	76.1	82.1	78.3
Happy	25.6	33.1	23.6	32.8	52.3	39.1	53.9	40.7
Sad	75.1	78.8	70.6	74.4	69.2	61.1	71.0	62.6
Excited	80.2	71.8	72.2	68.4	78.3	74.1	80.8	76.4
Frustrated	61.1	58.9	71.9	66.2	79.6	75.9	82.1	78.3

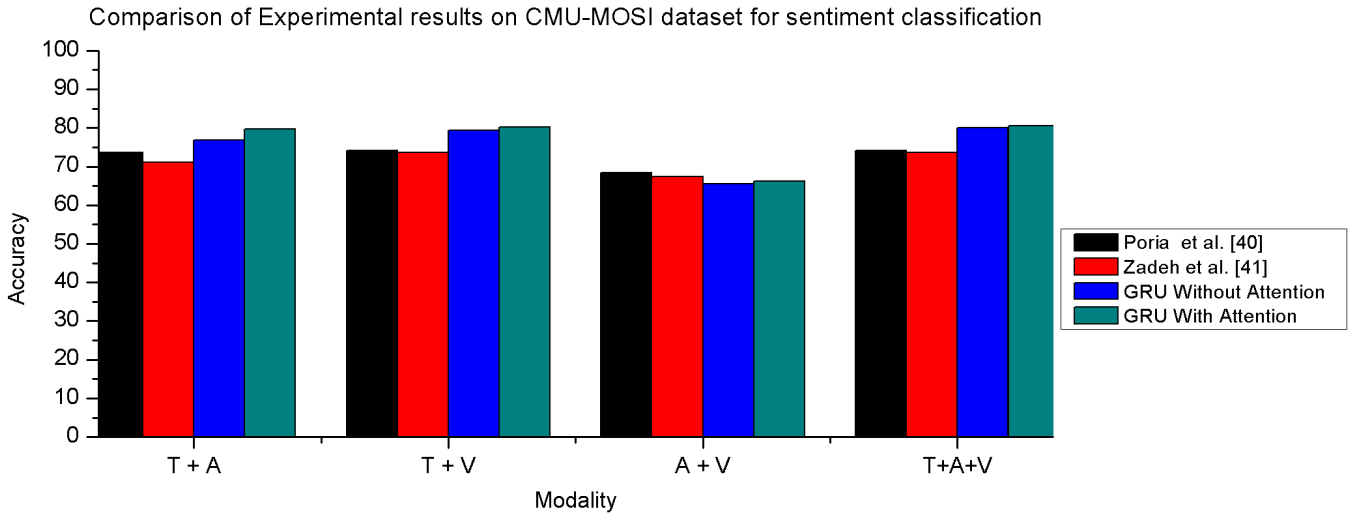

Fig. 6. Comparison of experimental results of the proposed method on CMU-MOSI dataset against standard baselines, Legend: T: Text, A: Audio, V: Video.

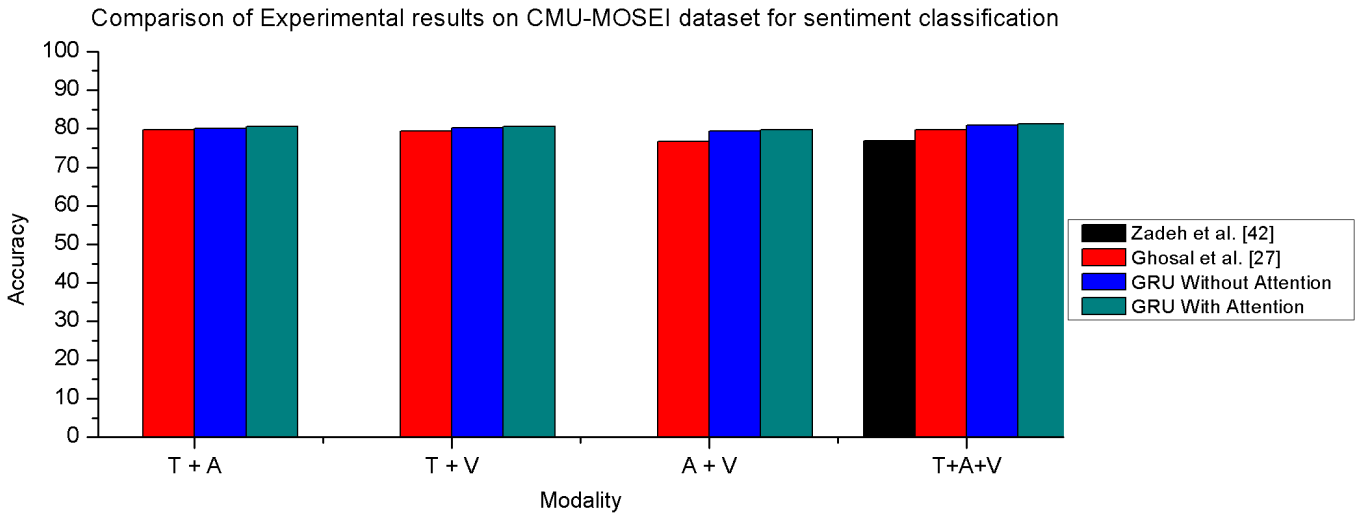

Fig. 7. Comparison of experimental results of the proposed method on CMU-MOSEI dataset against standard baselines.

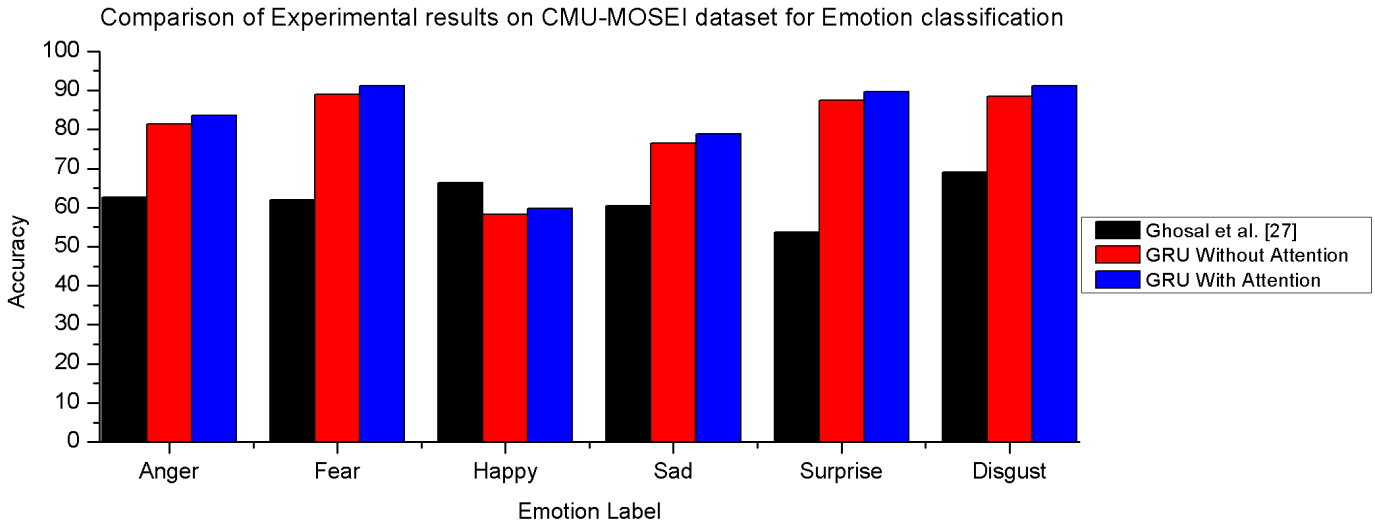

Fig. 8. Comparison of experimental results of the proposed method on CMU-MOSEI dataset against standard baselines.

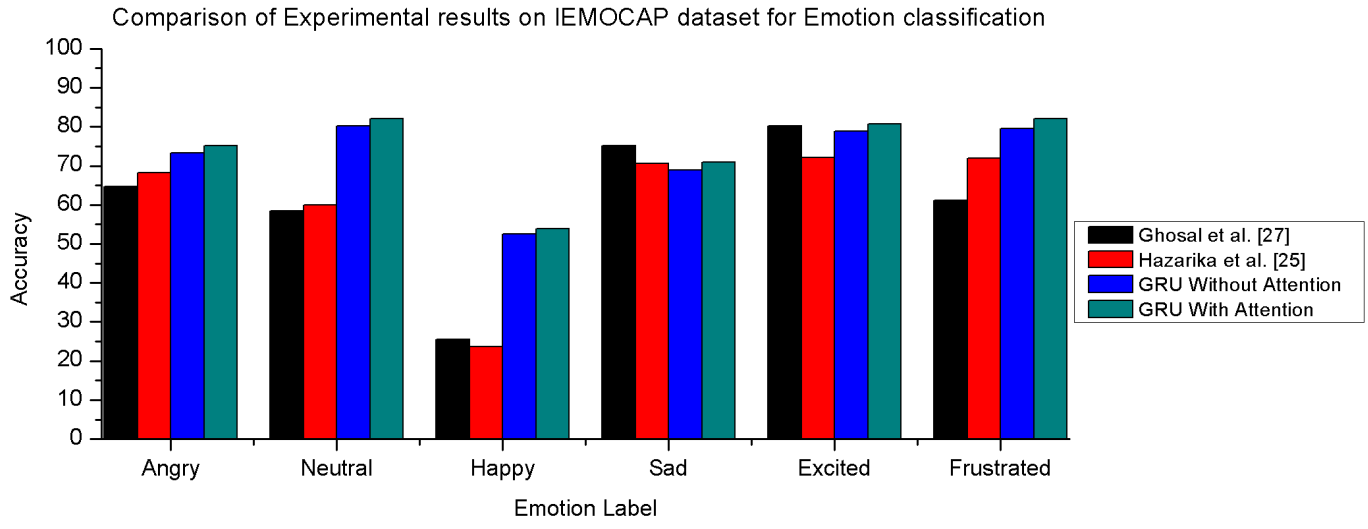

Fig. 9. Comparison of experimental results of the proposed method on IEMOCAP dataset against standard baselines

V. Conclusion and Future Work

The multimodal fusion, capturing interlocutor state of the participant, and understanding context between the utterances are the most important issues in multimodal sentiment analysis and emotion detection in conversation. In this paper first, features from individual modalities such as textual, acoustic, and visual features are extracted. Textual features are extracted using CNN and GloVe embedding's, audio features using open smile toolkit and visual features using 3D-CNN and facet toolkit. An attention-based pair-wise technique is used to extract the context between the utterances and understand the importance of constituent modalities before fusion. The recurrent neural network, more specifically gated recurrent Unit (GRU) based model is used to capture the interlocutor state and context extraction. By incorporating contextual information, the interlocutor state, and previous emotion state, the proposed model performs better than the standard baselines in terms of classification accuracy. In the future, we will explore techniques to address more than two participants in conversational videos. Also, we will study the feature selection methods to understand whether the emotion-specific features can improve the overall classification accuracy.

References

[1] S. Poria, E. Cambria and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2539–2544, 2015.

[2] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers), pp. 873-883, 2017.

[3] M. G. Huddar, S. S. Sannakki and V. S. Rajpurohit, "An Ensemble Approach to Utterance Level Multimodal Sentiment Analysis," in *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, Belgaum, India, pp. 145-150, 2018.

[4] M. G. Huddar, S. S. Sannakki and V. S. Rajpurohit, "Multi-level context extraction and attention-based contextual inter-modal fusion for multimodal sentiment analysis and emotion classification," International Journal of Multimedia Information Retrieval, vol. 9, no. 2, pp. 103-112, 2020, https://doi.org/10.1007/s13735-019-00185-8.

[5] M. G. Huddar, S. S. Sannakki and V. S. Rajpurohit, "Attention-based word-level contextual feature extraction and cross-modality fusion for sentiment analysis and emotion classification," *International Journal of Intelligent Engineering Informatics*, vol. 8, no. 1, pp. 1-18, 2020.

[6] S. Chen and Q. Jin, "Multi-modal conditional attention fusion for dimensional emotion prediction," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016.

[7] M. G. Huddar, S. S. Sannakki and V. S. Rajpurohit, "Multi-level feature optimization and multimodal contextual fusion for sentiment analysis and emotion classification," *Computational Intelligence*, vol. 36, no. 2, pp. 861-881, 2020.

[8] S. Poria, N. Majumder, R. Mihalcea and E. Hovy, "Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances," *arXiv preprint arXiv:1905.02947*, 2019.

[9] B. Kratzwald, S. Ilić, M. Kraus, S. Feuerriegel and H. Prendinger, "Deep learning for affective computing: Text-based emotion recognition in decision support," *Decision Support Systems*, vol. 115, pp. 24-35, 2018.

[10] M. G. Huddar, S. S. Sannakki and V. S. Rajpurohit, "A Survey of Computational Approaches and Challenges in Multimodal Sentiment Analysis," *International Journal of Computer Sciences and Engineering*, vol. 7, no. 1, pp. 876-883, 2019.

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

[12] V. Perez-Rosas, R. Mihalcea and L.-P. Morency, "Utterance-Level Multimodal Sentiment Analysis," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, 2013.

[13] J. G. Ellis, B. Jou and S.-F. Chang, "Why We Watch the News: A Dataset for Exploring Sentiment in Broadcast Video News," in *Proceedings of the 16th International Conference on Multimodal Interaction*, Istanbul, Turkey, 2014.

[14] F. Celli, B. Lepri, J.-I. Biel, D. Gatica-Perez, G. Riccardi and F. Pianesi, "The Workshop on Computational Personality Recognition 2014," in *Proceedings of the 22nd ACM international conference on Multimedia*, Orlando, Florida, USA , 2014.

[15] H. Kumar and B. Harish, "Automatic Irony Detection using Feature Fusion and Ensemble Classifier," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 7, pp. 70-79, 2019.

[16] H. Kumar, B. Harish and H. Darshan, "Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 5, pp. 109-114, 2019.

[17] M. G. Huddar, S. S. Sannakki and V. S. Rajpurohit, "Multimodal Emotion Recognition using Facial Expressions, Body Gestures, Speech, and Text Modalities," *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 8, no. 5, pp. 2453-2459, 2019.

[18] V. Rozgić, S. Ananthakrishnan, S. Saleem, R. Kumar and R. Prasad, "Ensemble of SVM trees for multimodal emotion recognition," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, Hollywood, CA, USA, 2013.

[19] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie and R. Cowie, "On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues," *Journal on Multimodal User Interfaces*, vol. 3, no. 1-2, p. 7–19, 2010.

[20] S. Poria, I. Chaturvedi, E. Cambria and A. Hussain, "Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis," in *IEEE 16th International Conference on Data Mining (ICDM)*, Barcelona, Spain, 2016.

[21] D. Datcu and L. J. Rothkrantz, "Semantic audio-visual data fusion for automatic emotion recognition," *Emotion recognition: a pattern analysis approach*, pp. 411-435, 2014.

[22] J. Chung, C. Gulcehre, K. Cho and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," *arXiv:1412.3555*, 2014.

[23] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh and E. Cambria, "DialogueRNN: An Attentive RNN for Emotion Detection in Conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[24] A. Zadeh, P. P. Liang, S. Poria, E. Cambria and L.-P. Morency, "Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph," in *Proceedings of the 56th Annual Meeting of the As-sociation for Computational Linguistics (Volume 1: Long Papers)*, 2018.

[25] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria and R. Zimmermann, "ICON: interactive conversational memory network for multimodal emotion detection," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[26] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, Louis-Philippe Morency and R. Zimmermann, "Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long Papers), New Orleans, Louisiana, 2018.

[27] D. Ghosal, M. S. Akhtar, D. Chauhan, S. Poria, A. Ekbal and P. Bhattacharyya, "Contextual inter-modal attention for multi-modal sentiment analysis," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[28] M. Wan, G. Yang, S. Gai and Z. Yang, "Two-dimensional discriminant locality preserving projections (2DDLPP) and its application to feature extraction via fuzzy set," *Multimedia tools and applications*, vol. 76, no. 1, pp. 355-371, 2017.

[29] M. Wan, M. Li, G. Yang, S. Gai and Z. Jin, "Feature extraction using two-dimensional maximum embedding difference," *Information Sciences*, vol. 274, pp. 55-69, 2014.

[30] M. Wan, Z. Lai, G. Yang, Z. Yang, F. Zhang and H. Zheng, "Local graph embedding based on maximum margin criterion via fuzzy set," *Fuzzy Sets and Systems*, vol. 318, pp. 120-131, 2017.

[31] M. Magdin, T. Sulka, J. Tomanová and M. Vozár, "Voice Analysis Using PRAAT Software and Classification of User Emotional State," *IJIMAI*, vol. 5, no. 6, pp. 33-42, 2019.

[32] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335-359, 2008.

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

[34] F. Eyben, M. Wöllmer and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, Barcelona, Spain, 2013.

[35] A. Karpathy, G. Toderici, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-scale Video Classification with Convolutional Neural Networks," in *Proceedings of International Computer Vision and Pattern Recognition*, 2014.

[36] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv:1301.3781*, 2013.

[37] Y. W. Teh and G. E. Hinton, "Rate-coded restricted Boltzmann machines for face recognition," in *Proceedings of the 13th International Conference on Neural Information Processing Systems*, Cambridge, MA, USA, 2000.

[38] S. Ji, W. Xu, M. Yang and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221 - 231, 2013.

[39] D. a. B. J. Kingma, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*, vol. 15, 2014.

[40] S. Poria, I. Chaturvedi, E. Cambria and A. Hussain, "Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis," in *2016 IEEE 16th international conference on data mining (ICDM)*, Barcelona, Spain, 2016.

[41] A. Zadeh, M. Chen, S. Poria, E. Cambria and L.-P. Morency, "Tensor Fusion Network for Multimodal Sentiment Analysis," in *Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 2017.

[42] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), Melbourne, Australia, 2018.

Mahesh G Huddar

Mahesh G. Huddar is working as an Assistant Professor in Computer Science and Engineering at Hirasugar Institute of Technology, Nidasoshi, Belagavi, India and he is currently pursuing Ph.D. studies at the Visvesvaraya Technological University, Belagavi, India in the Department of Computer Science and Engineering. He received his Master and Bachelor of Science degrees from the Visvesvaraya Technological University, Belagavi, India in 2014 and 2008, respectively. He has published a many papers in journals, International, and National conferences. His main research interests include Machine Learning, Deep Learning, Multimodal Sentiment Analysis, and Multimodal Emotion Detection. He is a member of the IEEE.

Sanjeev S Sannakki

Prof. Sanjeev S. Sannakki is working as a Professor in the Department of Computer Science and Engineering at Gogte Institute of Technology, Belgaum, Karnataka, India. He pursued his B.E. in Electronics & Communication from Karnataka University Dharwad in 2004, M.Tech, and Ph.D. from VTU, Belagavi in 2009, and 2016 respectively. His research areas include Image Processing, Cloud Computing, Computer Networks, and Data Analytics. He has published a good number of papers in journals, International, and National conferences. He is the reviewer for a few international journals and conferences. He is also a life member of CSI, and ISTE associations.

Vijay S Rajpurohit

Prof. Vijay S Rajpurohit is working as a Professor in the Department of Computer Science and Engineering at Gogte Institute of Technology, Belgaum, Karnataka, India. He pursued his B.E. in Computer Science and Engineering from Karnataka University Dharwad, M.Tech from N.I.T.K Surathkal, and Ph.D. from Manipal University, Manipal in 2009. His research areas include Image Processing, Cloud Computing, and Data Analytics. He has published a good number of papers in journals, International, and National conferences. He is the reviewer for a few international journals and conferences. He is the associate editor for two international journals and a Senior Member of the International Association of CS and IT. He is also the life member of SSI, ISC, and ISTE associations.