International Journal of Multimedia Information Retrieval (2020) 9:103–112
https://doi.org/10.1007/s13735-019-00185-8

REGULAR PAPER

Multi-level context extraction and attention-based contextual inter-modal fusion for multimodal sentiment analysis and emotion classification

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

Received: 30 April 2019 / Revised: 17 November 2019 / Accepted: 25 November 2019 / Published online: 14 December 2019

Abstract
The recent advancements in the Internet technology and its associated services, led the users to post a large amount of multimodal data into social media Web sites, online shopping portals, video repositories, etc. The availability of the huge amount of multimodal content, multimodal sentiment classification, and affective computing has become the most researched topic. The extraction of context among the neighboring utterances and considering the importance of inter-modal utterances before multimodal fusion are the most important research issues in this field. This article presents a novel approach to extract the context at multiple levels and to understand the importance of inter-modal utterances in sentiment and emotion classification. Experiments are conducted on two publically accepted datasets such as CMU-MOSI for sentiment analysis and IEMOCAP for emotion classification. By incorporating the utterance-level contextual information and importance of inter-modal utterances, the proposed model outperforms the standard baselines by over 3% in classification accuracy.

Keywords Attention model · Inter-modal fusion · Multi-level contextual information · Bidirectional recurrent neural network

1 Introduction

The recent growth in World Wide Web (WWW), social media networking platforms, online shopping portals, travel guide Web sites, and video repositories led the user to post their opinion about a product or an entity in textual–audio–visual format or in a combination of two or more modalities at a time [1]. Many organizations and online portals are using this information to understand the people's opinions and emotions toward their service or quality of a product or analyze customer complaints, etc. The information extracted from audio-visual content helps to improve the quality of life of a customer by allowing a customer to choose the best product to buy, select the best available service or to decide which movie to watch or to select the best tourist destination, etc.

The traditional sentiment analysis and emotion detection were restricted to textual modality. In contrast, more recently multimodal sentiment analysis and affective computing have gained a lot of interest among the researchers. The primary advantage of analyzing multimodal content over unimodal content is to leverage the variety of information from multiple modalities to build a more accurate system in terms of classification accuracy [1]. The availability of facial expressions in the visual modality and vocal cues in audio modality creates a more robust system and improves the accuracy of the overall system compared to the system with only textual modality [2].

Multimodal content on the Internet has three modalities such as audio, visual, and text that is a transcribed audio. In traditional multimodal data analysis, features are extracted from individual modalities and use either early feature fusion (feature concatenation) [3] or late fusion [4] to fuse the features

✉ Mahesh G. Huddar
mailtomgh1@gmail.com

Sanjeev S. Sannakki
sannakkisanjeev@gmail.com

Vijay S. Rajpurohit
vijaysr2k@yahoo.com

[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

extracted from individual modalities. Usually, in multimodal sentiment analysis and emotion detection, the video is divided into small opinion segments, also called utterances. Each utterance is manually labeled with sentiment polarity or emotion label. In utterance-level sentiment analysis, each utterance is analyzed individually irrespective of neighboring utterances [5, 6]. As a video is a sequence of opinion segments, neighboring utterances influence (contextual information) the outcome of current utterance. Utterance-level sentiment analysis and traditional fusion techniques cannot extract context from multiple utterances. Also, in multimodal sentiment analysis with context modeling, the attention is applied directly to the contextual feature vector. This approach considers only the modality-specific contextual features and restricts the influence of other modalities. These issues are addressed in the proposed model by incorporating the bidirectional recurrent neural network-based models to extract the contextual features at different levels such as at unimodal, bimodal, and trimodal levels. Also, a correlation among the multiple modalities at the utterance level is computed by considering the target utterance context and the context of other utterances. This helps to understand the importance of utterance from different modalities before fusion.

The main contributions of the proposed work are as follows:

- An effective attentive model is proposed to accommodate the features from multiple modalities as well as neighboring utterances in multimodal sentiment analysis and affective computing.
- The model demonstrates that better multimodal feature representation can be generated by combining attentive inter-modality features and contextual utterances within the modality.
- The state-of-the-art model is presented for sentiment and emotion classification and tested on two different standard datasets.

The remaining paper is structured as follows: Recent related work in multimodal sentiment analysis and emotion detection is discussed in Sect. 2. The proposed multi-level context extraction and attentive contextual inter-modality fusion are discussed in Sect. 3. The experimental results on two different standard datasets are presented and discussed in Sect. 4. Finally, Sect. 5 presents the future work in multimodal sentiment analysis and emotion detection and concludes the paper.

2 Related work

Due to the enormous growth in social media, the availability of affordable Internet facilities and smartphones led the user to post reviews in different forms like text, image, video, audio, etc. The availability of multimodal data creates curiosity among the researchers to leverage the information from multiple modalities to build a more accurate system. The data from online social media networks, textual blogs, video blogs, shopping portals, and travel guide Web sites are analyzed to predict the user sentiment toward a service, a product, and an entity [7]. Traditionally, sentiment and emotion classification is performed on textual data [8], but recently, multimodal content such as text–audio–visual content is being used in sentiment and emotion classification [9]. Most of the work in textual sentiment analysis is performed on the English language, but a more recently multilingual issue is considered by many researchers [10, 11]. Unimodal or multimodal sentiment classification methods are designed using either sentiment lexicons [12] or ontology-based features [13] or handcrafted features [14] or sentiment networks [15].

It seems like sentiment analysis and emotion classification are simple classification task, but in actual scenario, it is a very challenging field of research, as it is required to consider and accommodate many tasks such as topic detection, analysis and tracking [16], word polarity analysis [17], text summarization [18], aspect consideration [19], and subjectivity recognition and analysis [20]. The business organizations are using sentiment analysis and emotion classification to address many real-world issues such as forecasting political results [21], improve quality of service and product [22], improvement in financial sector [23], movie box office prediction [24], in e-tourism [25], in e-health [26], etc.

The audio–visual features are concatenated to form the bimodal feature vector [27], and the results show that the proposed bimodal emotion detection system achieves better accuracy than the unimodal-based systems. Later researchers used either early feature fusion (feature concatenation) [3] or late fusion [4] or hybrid fusion [28] to fuse the features extracted from individual modalities to create bimodal and trimodal audio–visual sentiment analysis and emotion detection models [29, 30].

In Rozgić et al. [31], lexicon, acoustic, and visual features are extracted and the authors first use the SVM classifier on individual modalities. Finally, the ensemble approach is used to find the effective label for the utterance [32]. Also, they demonstrated that their proposed ensemble model achieves better classification accuracy than the traditional methods. In Eyben et al. [33], linguistic and acoustic features are extracted using the 3D activation valance method for emotion classification. Using convolutional neural networks, textual–visual–acoustic features are extracted and multiple kernel learning is used as a classifier for analyzing sentiment and emotion [34]. More recently, memory fusion and tensor fusion techniques are proposed for sentiment analysis and emotion classification [35]. The proposed method considers the impact of neighboring utterances to leverage the utterance-level contextual information and attention-based

Table 1 Train–test distribution of IEMOCAP dataset

	Happy	Angry	Sad	Neutral
Train	1194	933	839	1324
Test	433	157	238	380

inter-modality fusion to accommodate the importance of modality before fusion.

3 Proposed methodology

This section discusses the proposed multi-level context extraction and attentive inter-modal fusion technique in detail. The brief overview of the proposed methodology is as follows:

- First unimodal features (acoustic, textual, and visual) are extracted at utterance level.
- Unimodal contextual features are extracted from each modality using bidirectional recurrent neural network-based models.
- Multimodal attention, context extraction, and inter-modal fusion are performed.
- Finally, a trimodal attentive contextual feature vector is used for sentiment analysis and emotion detection.

3.1 Dataset used

The proposed methodology is demonstrated on two publically available and most researched standard datasets such as CMU-MOSI [36] and IEMOCAP [37] for multimodal sentiment analysis and emotion classification, respectively.

3.1.1 IEMOCAP

The IEMOCAP dataset is an acted two-way dyadic conversational dataset among multiple speakers. Each conversational video is divided into multiple small opinion segments called utterances. Multiple annotators manually assign emotion labels to each utterance. The average voted result of annotators is taken as a final label for the utterance. To be in-line with the recent research work in multimodal affective computing, utterances with four labels such as happiness (excitement), sadness, anger and neutral are considered for the study. The train–test distribution of the IEMOCAP dataset is shown in Table 1.

3.1.2 CMU-MOSI

The CMU-MOSI dataset is prepared by collecting videos from YouTube. The dataset contains a totally of 93 videos, with 89 distinct speakers reviewing various products. Each video review is divided into multiple opinion segments called utterances. There are 2199 opinion segments or utterances in the CMU-MOSI dataset. The average length of each segment or utterance is 4.2 s, with each having an average of 12 words. Five assessors manually assign a label to each of the utterances in the range of -3 to $+3$. -3 and $+3$ indicate strongly negative and strongly positive polarities, respectively. The voted average is taken as the final label for the utterance. To be in-line with the recent research work in multimodal sentiment analysis, utterances with only positive and negative polarities are considered for the study. The train–test distribution of the CMU-MOSI dataset is given in Table 2.

Table 2 Train–test distribution of CMU-MOSI dataset

	Positive	Negative
Train	709	738
Test	467	285

3.2 Feature extraction

The methodology used to extract features from multiple modalities such as audio–text–visual is described in this section.

3.2.1 Audio feature extraction

OpenSMILE [38] toolkit is an open-source software used for acoustic feature extraction and analysis. Before extracting acoustic features from an audio modality, voice normalization and voice intensity thresholding are performed using Z-score standardization [39]. A total of 6392 features such as MFCC, voice intensity, pitch, amplitude mean, root quadratic mean, arithmetic mean, standard deviation, quartile ranges, interquartile ranges, and standard deviation are extracted from each utterance using OpenSMILE toolkit at a frame rate of 30 Hz with a sliding window of size 100 ms.

Let A_i be the audio feature vector of ith utterance, then the audio feature vector A is defined as follows:

$$A = \langle A_1, A_2, A_3, \ldots, A_n \rangle \tag{1}$$

where n is the number of utterances.

3.2.2 Textual feature extraction

First, each utterance is manually transcribed to get the textual data. More recently, many researchers have used convolutional neural networks (CNN) [40] and proved their efficacy in natural language processing. The word-level textual representation is prepared using Word2vec vectors [41]. Three

convolutional layers are used to process these Word2vec representations. Each layer uses 4, 3, and 2 filters of size 50, 75, and 100 feature maps, respectively. After every convolutional layer, 2X2 sized max-pooling is used. The output of the convolutional layer is fed to a fully connected dense layer with 600 computing neurons, followed by a softmax classifier. Relu [42] is used as an activation function. The output of the softmax classifier is used as the textual feature vector.

Let T_i be the textual feature vector of ith utterance, then the textual feature vector T is defined as follows:

$$T = \langle T_1, T_2, T_3, \ldots, T_n \rangle \tag{2}$$

3.2.3 Visual feature extraction

The complex tasks such as object detection, recognition, tracking, and classification are successfully addressed using 3D convolutional neural networks [43]. The results presented in Ji et al. [43] motivate to use 3D CNN model. Temporal features across the frames and frame-level features are extracted using 3D CNN model in Poria et al. [6]. The same procedure is adopted to extract visual features.

Let V_i be the visual feature vector of ith utterance, then the visual feature vector V is defined as follows:

$$V = \langle V_1, V_2, V_3, \ldots, V_n \rangle \tag{3}$$

3.3 Multi-level context extraction and attentive contextual inter-modal fusion

The context extraction between the neighboring utterances in a video and multimodal fusion are the primary requirements of multimodal sentiment analysis and emotion classification. This section describes the methodology followed in context extraction at multiple levels and the attentive contextual inter-modal fusion procedure to merge the information extracted from multiple modalities.

3.3.1 Problem with traditional fusion techniques

Most of the recent research work in the field of multimodal sentiment and emotion classification uses either early feature fusion (feature concatenation) [3] or late fusion [4] techniques. In early fusion, the feature vectors from multiple modalities are appended to form the final feature vector. Classification is performed on this resultant feature vector. In late or decision-level fusion, initially classification is performed on modality-specific feature vectors and then, the classification result from each modality is combined using ensemble techniques to get the final classification label. Figure 1 shows the traditional fusion techniques such as early and late fusion techniques. These simplistic traditional fusion techniques cannot extract the contextual information between the neighboring utterances or redundant and complex affective information from different modalities such as audio, video, and text. Also, every modality and utterance may not have the same importance in the sentiment and emotion classification. These issues are addressed in the proposed methodology by extracting the context between the neighboring utterances at multiple levels by using the bidirectional recurrent neural network-based models, and attention-based model is used to calculate the importance of each modality and utterance before fusion.

3.3.2 Unimodal context extraction

Each video is a sequence of segments or utterances, and each utterance may not be a complete sentence. Hence, the outcome of a segment or utterance may depend on the outcome of neighboring utterance. Hence, the outcome of preceding utterance and succeeding utterance must be considered while

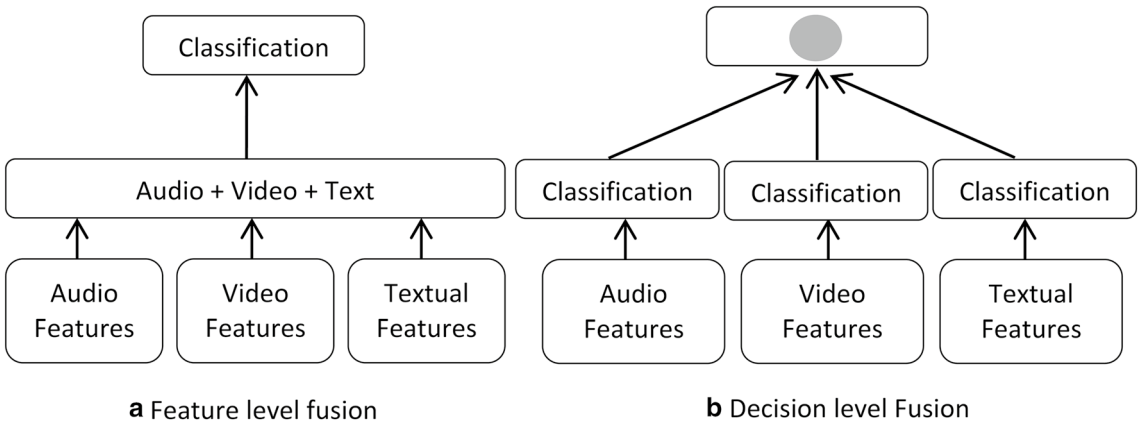

Fig. 1 Traditional fusion techniques **a** feature level or early fusion: the concatenation of features from multiple modalities, **b** decision level or late fusion: a fusion of predictions from multiple modalities

extracting the sentiment polarity and affective label of an utterance. The bidirectional recurrent neural network-based model is used to extract the contextual data between the neighboring utterances before fusion.

Let A_{ij}, T_{ij} and V_{ij} be the audio, textual, and visual feature vector of ith utterance of jth video and N be the number of utterances in a video, then the contextual audio, textual, and visual feature vectors of jth video A_{cj}, T_{cj} and V_{cj}, respectively, are defined as follows:

$$A_{cj} = \text{Context\_Extract}\left(A_{1j}, A_{2j}, A_{3j}, \ldots, A_{Nj}\right) \tag{4}$$

$$T_{cj} = \text{Context\_Extract}\left(T_{1j}, T_{2j}, T_{3j}, \ldots, T_{Nj}\right) \tag{5}$$

$$V_{cj} = \text{Context\_Extract}\left(V_{1j}, V_{2j}, V_{3j}, \ldots, V_{Nj}\right) \tag{6}$$

3.3.3 Multimodal attention and context extraction

Assume a particular video V_j has n opinion segments or utterances with A_{cj}, T_{cj} and V_{cj} being the contextual audio, textual, and visual feature vectors. Pair-wise attention between the modalities such as (AT, AV, TV) is performed to obtain the bimodal representation. The bimodal context is extracted using a bidirectional recurrent neural network-based model. Bimodal attention matrix representations are concatenated to obtain a trimodal attention matrix followed by trimodal context extraction is performed. The trimodal contextual feature vector is fed to the softmax classifier to obtain the final classification label for the utterance.

Modality-specific contextual feature vectors A_{cj}, T_{cj} and V_{cj} contain contextual information between the utterances of each modality. Pair-wise matching matrices $M_1(AT), M_2(TA), M_1(AV), M_2(VA), M_1(TV)$ and $M_2(VT)$ are computed for each utterance over two representations that account for the cross-modality information.

$$M_1(AT) = A_{cj} \cdot T_{cj}^T \text{ and } M_2(TA) = T_{cj} \cdot A_{cj}^T \tag{7}$$

$$M_1(AV) = A_{cj} \cdot V_{cj}^T \text{ and } M_2(VA) = V_{cj} \cdot A_{cj}^T \tag{8}$$

$$M_1(TV) = T_{cj} \cdot V_{cj}^T \text{ and } M_2(VT) = V_{cj} \cdot T_{cj}^T \tag{9}$$

A softmax function is used to compute the probability distribution scores $W_1(AT), W_2(TA), W_1(AV), W_2(VA), W_1(TV)$ and $W_2(VT)$ over each utterance or segment of bimodal matching matrices M_1 and M_2. This computes the contextual attention weights for each of the utterances in a particular video. Modality-specific attention weights are computed using soft attention over the bimodal contextual attention.

The probability distribution scores $W_1(AT), W_2(TA)$ of an audio–text combination of modality are calculated as follows:

$$W_1(AT)_i^j = \frac{e^{M_1(AT)_i^j}}{\sum_{k=1}^{n} e^{M_1(AT)_i^k}} \quad i, j = 1, 2, 3, \ldots, n \tag{10}$$

$$W_2(TA)_i^j = \frac{e^{M_2(TA)_i^j}}{\sum_{k=1}^{n} e^{M_2(TA)_i^k}} \quad i, j = 1, 2, 3, \ldots, n \tag{11}$$

Similarly, probability distribution scores $W_1(AV), W_2(VA), W_1(TV)$ and $W_2(VT)$ of audio–video and text–video combination of modalities are calculated.

The modality-specific attention weights are calculated as follows:

$$W_1(T) = W_1(AT) \cdot T \text{ and } W_2(A) = W_2(TA) \cdot A \tag{12}$$

$$W_1(V) = W_1(AV) \cdot V \text{ and } W_2(A) = W_2(VA) \cdot A \tag{13}$$

$$W_1(V) = W_1(TV) \cdot V \text{ and } W_2(T) = W_2(VT) \cdot T \tag{14}$$

Bimodal attention matrices are calculated by element-wise matrix multiplication of modality-specific weights and other modality feature vector.

$$A_1(AT) = W_1(T) \cdot A \text{ and } A_2(TA) = W_2(A) \cdot T \tag{15}$$

$$A_1(AV) = W_1(V) \cdot A \text{ and } A_2(VA) = W_2(A) \cdot V \tag{16}$$

$$A_1(TV) = W_1(V) \cdot T \text{ and } W_2(VT) = W_2(T) \cdot V \tag{17}$$

Bimodal attention matrices A_1 and A_2 of a combination of modalities are concatenated to get the bimodal feature vector AT, AV and TV.

$$AT = \text{Concatenate}\left(A_1(AT), A_2(TA)\right) \tag{18}$$

$$AV = \text{Concatenate}\left(A_1(AV), A_2(VA)\right) \tag{19}$$

$$TV = \text{Concatenate}\left(A_1(TV), A_2(VT)\right) \tag{20}$$

Next, the bidirectional recurrent neural network-based model is used to extract a bimodal contextual feature vector from a bimodal feature vector. Bimodal contextual feature vectors are concatenated using residual skip attention-based networks to obtain a trimodal attention matrix. This contextual trimodal attention matrix is used for final classification.

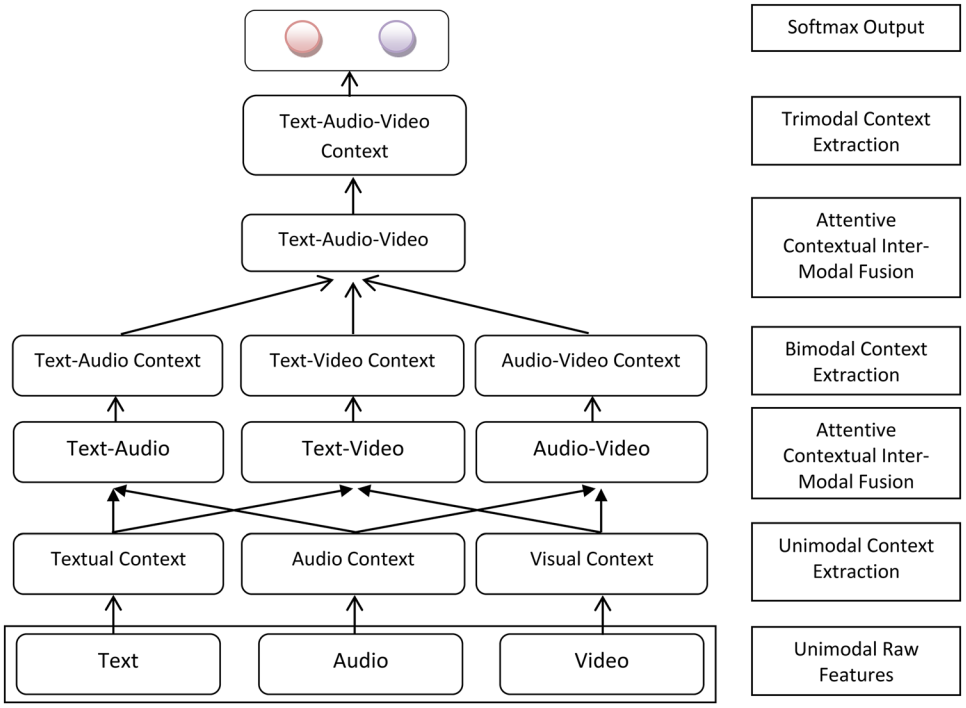

Fig. 2 Proposed multi-level context extraction and attention-based contextual inter-modal fusion

3.4 Classification and training

The contextual trimodal attention matrix f is fed as input to softmax classifier to predict the sentiment polarity or affective label $\hat{y}$ of the testing utterance U. The softmax classifier is defined as follows:

$$p(y|U) = \text{softmax}\left(w^{(s)}f + b^{(s)}\right) \tag{21}$$

$$\hat{y} = \arg\max_{y} p(y|U) \tag{22}$$

where $b^{(s)}$ and $w^{(s)}$ are bias and weight matrices, respectively, p is the class probability, and $\hat{y}$ is the final label for the testing utterance.

Given the training data, the proposed model was trained, by employing the cross-entropy loss function $L(\theta)$. The cross-entropy loss function is defined as follows:

$$L(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_i^j \log \hat{y}_i^j + \lambda \sum_{k=0}^{N} \theta_k^2 \tag{23}$$

where N the number of opinion segments or utterances in training dataset, and y_s the true sentiment or emotion class of the sth utterance. $\hat{y}_s$ is the predicted sentiment or emotion class of the sth utterance, and M the number of sentiment or emotion classes. Θ is the parameter set term, and $L2$ regularization term.

The proposed attention-based multi-level context extraction and attention-based contextual inter-modal fusion for multimodal sentiment analysis and emotion classification are shown in Fig. 2, and Table 3 summarizes our proposed methodology.

4 Result analysis and discussion

The proposed model is implemented using Python-based Keras library with Tensor Flow as backend. Tesla K80 graphical processor unit with 12 GB of RAM is used to conduct the experiments. The results of the proposed model are compared with the two baselines [34, 35]. Convolutional neural network-based model is used to extract textual features; visual and acoustic features are extracted using CLM-Z and OpenSMILE toolkits, respectively, in Poria et al. [34]. Unimodal features are concatenated to build the bimodal and trimodal feature vectors. Novel feature fusion model called tensor fusion is proposed in Zadeh et al. [35].

The performance of the proposed model is compared with the two standard baselines in Tables 4 and 5 for emotion classification and sentiment polarity on IEMOCAP and CMU-MOSI dataset, respectively. The impact of recurrent neural units such as gated recurrent unit (GRU), bidirectional GRU (biGRU), bidirectional GRU with attention, long short-term memory (LSTM), bidirectional LSTM (biLSTM), and bidirectional LSTM with attention on the proposed method is analyzed. The results show that the proposed model performs better in terms of classification accuracy compared to the baselines in all combinations of modalities except the audio–video combination of modalities. Also, note that the LSTM-based recurrent unit is performing better than

Table 3 Algorithm: multi-level context extraction and attention-based contextual inter-modal fusion

Algorithm	
1: $Procedure\ FeatureExtraction(U)$	$Procedure\ to\ extract\ unimodal\ features$
2: $\quad for\ i\ in\ 1\ to\ N\ do:$	
3: $\quad\quad A_i \longleftarrow audiofeatures(U_i)$	
4: $\quad\quad T_i \longleftarrow textualfeatures(U_i)$	
5: $\quad\quad V_i \longleftarrow visualfeatures(U_i)$	
6: $Procedure\ Context\_Extract(f)$	$Procedure\ to\ extract\ unimodal\ context$
7: $\quad f_c \longleftarrow biLSTM(f)$	
8: $\quad return\ (f_c)$	
9: $Procedure\ BimodalFusion(X, Y)$	$Procedure\ for\ Bimodal\ fusion$
10: $\quad For\ j\ in\ 1\ to\ N\ do:$	
11: $\quad\quad M_1(XY) = X_{cj}.Y_{cj}^{\mathrm{T}}$	
12: $\quad\quad M_2(YX) = Y_{cj}.X_{cj}^{\mathrm{T}}$	
13: $W_1(XY)_i^j = \frac{e^{M_1(XY)_i^j}}{\sum_{k=1}^{n} e^{M_1(XY)_i^k}}\ i, j = 1, 2, 3, \ldots \ldots n$	
14: $W_2(YX)_i^j = \frac{e^{M_2(YX)_i^j}}{\sum_{k=1}^{n} e^{M_2(YX)_i^k}}\ i, j = 1, 2, 3, \ldots \ldots n$	
15: $W_1(X) = W_1(XY)\ .\ Y$	
16: $W_2(Y) = W_2(YX)\ .\ X$	
17: $A_1(XY) = W_1(Y)\ .\ X$	
18: $A_2(YX) = W_2(X)\ .\ Y$	
19: $XY = Concatenate\ (A_1(XY), A_2(YX))$	
20: $retrun\ XY$	
21: $Procedure\ TrimodalFusion(AT, AV, TV)$	$Procedure\ for\ Trimodal\ fusion$
22: $\quad return\ \ Concataenate\ (AT, AV, TV)$	
23: $Procedure\ Classification\ (f, U)$	$Procedure\ for\ classification\ of\ utterance\ \ into\ discrete\ number\ of\ classes$
24: $\quad for\ i\ in\ 1\ to\ N\ do:$	
25: $\quad\quad p\ (y\|U)\ =\ softmax\ (\ w^{(s)}f\ +\ b^{(s)}\ )$	
26: $\quad \hat{y}\ =\ \underset{y}{argmax}\ p\ (y\|U)$	
27: $\quad return\ (\hat{y})$	
28: $FeatureExtraction(U)$	$Unimodal\ Feature\ Extraction$
29: $for\ f \in \{A, T, V\}$	$Unimodal\ Context\ Extraction$
30: $\quad f_c \longleftarrow\ Context\_Extract(f)$	
31: $AT \longleftarrow BimodalFusion(A_c, T_c)$	$Bimodal\ Fusion$
32: $AV \longleftarrow BimodalFusion(A_c, V_c)$	
33: $TV \longleftarrow BimodalFusion(T_c, V_c)$	
34: $for\ f \in \{AT, AV, TV\}$	$Bimodal\ Context\ Extraction$
35: $\quad f_c \longleftarrow Context\_Extract(f)$	
36: $AVT \longleftarrow TrimodalFusion(AT_c, AV_c, AV_c)$	$Trimodal\ Fusion$
37: $AVT_c \longleftarrow Context\_Extract(AVT)$	$Trimodal\ Context\ Extraction$
38: $C \longleftarrow Classification\ (AVT_c, U)$	$Classification$

Table 4 Performance of the proposed model compared with baselines for affective classification using IEMOCAP dataset legend: *T* text, *A* audio, *V* video

Modality	Poria et al. [34]	Zadeh et al. [35]	Multimodal fusion without contextual features		Multimodal fusion with contextual features		Attention-based intermodal fusion with contextual features	
			GRU	LSTM	biGRU	biLSTM	biGRU with attention	biLSTM with attention
T+*A*	73.7	71.1	74.29	76.18	74.48	77.14	77.28	78.85
T+*V*	74.1	73.7	74.15	75.94	74.23	76.37	77.17	78.32
A+*V*	68.4	67.4	65.13	66.21	65.46	66.91	66.97	67.75
T+*A*+*V*	74.1	73.6	76.87	77.98	76.93	79.52	78.87	80.87

Table 5 Performance of the proposed model compared with baselines for sentiment classification using CMU-MOSI dataset

Modality	Poria et al. [34]	Zadeh et al. [35]	Multimodal fusion without contextual features		Multimodal fusion with contextual features		Attention-based intermodal fusion with contextual features	
			GRU	LSTM	biGRU	biLSTM	biGRU with attention	biLSTM with attention
T+*A*	77.3	77.0	76.13	78.54	77.11	79.37	78.38	79.82
T+*V*	77.8	77.1	75.83	78.17	76.91	79.14	78.15	79.86
A+*V*	57.9	56.5	57.17	57.14	57.34	58.54	58.37	59.32
T+*A*+*V*	78.7	77.2	76.73	78.62	77.78	79.78	78.85	80.33

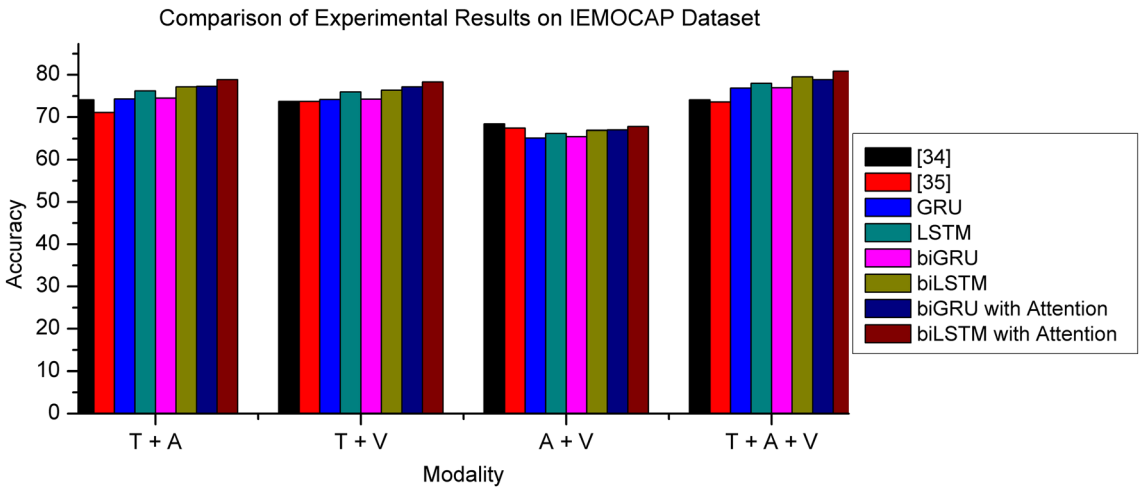

Fig. 3 Comparison of experimental results on IEMOCAP dataset

GRU-based recurrent units in both sentiment and emotion classifications. The results show that the performance of a model with contextual features is better than the model without contextual features. Also, attentive contextual intermodal fusion accommodates both the contextual features and the importance of inter-modal utterance, and it performs better than the other two models.

Figure 3 shows the comparison of experimental results on the IEMOCAP dataset for emotion classification.

Figure 4 shows the comparison of experimental results on the CMU-MOSI dataset for sentiment classification.

5 Conclusion and future work

The main issues in multimodal sentiment analysis and emotion detection are contextual information extraction and multimodal fusion. This paper presents a multi-level contextual

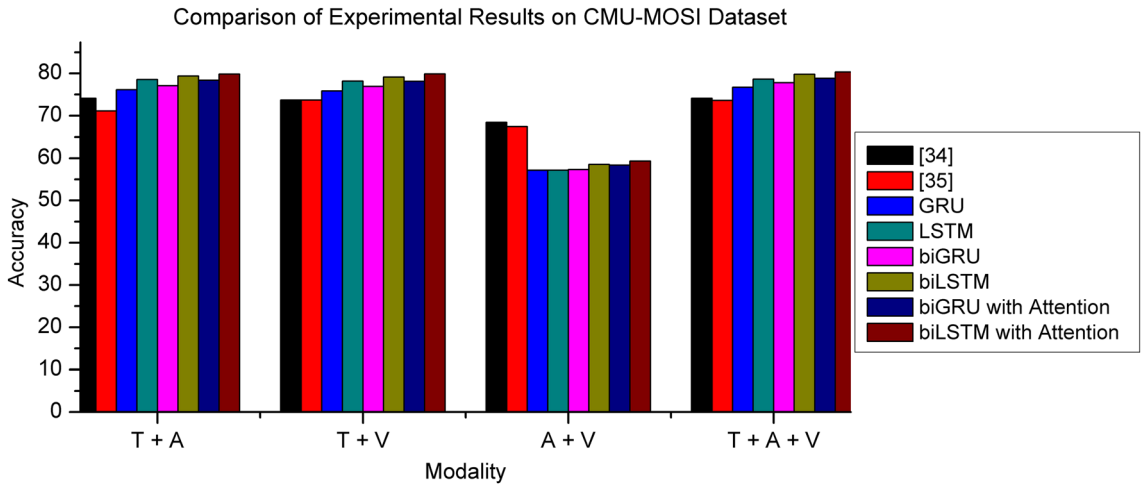

Fig. 4 Comparison of experimental results on CMU-MOSI dataset

feature extraction technique using a bidirectional recurrent neural network-based model. Also, the contribution of each modality is different in sentiment and emotion classification at the utterance level. Hence, the proposed model proposes the attention-based inter-modality fusion for accommodating the importance of each inter-modal utterance at multimodal fusion. Contextual attentive unimodal features are merged two–two at a time to get bimodal features, and all are merged to get trimodal feature vectors. After every level of fusion, contextual features are extracted. Finally, the experimental result on CMU-MOSI and IEMOCAP dataset shows that the trimodal combination and bimodal combination of modalities perform better than the unimodal modalities. Also, the proposed model outperforms the standard baselines by a huge margin. The future work includes the investigation of the impact of a subset of features and class-specific features on classification accuracy. Also, we explore the methods to improve the quality of unimodal features, which will further improve the classification accuracy of the overall multimodal system.

References

1. Poria S, Cambria E, Bajpai R, Hussain A (2017) A review of affective computing: from unimodal analysis to multimodal fusion. Inf Fusion 37:98–125
2. Huddar MG, Sannakki SS, Rajpurohit VS (2019) Multimodal emotion recognition using facial expressions, body gestures, speech, and text modalities. Int J Eng Adv Technol (IJEAT) 8(5):2453–2459
3. Rosas VP, Mihalcea R, Morency L-P (2013) Multimodal sentiment analysis of Spanish online. IEEE Intell Syst 28(3):38–45
4. Ellis JG, Jou B, Chang S-F (2014) 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
5. Poria S, Cambria E, Gelbukh A (2015) Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In: EMNLP, pp 2539–2544
6. Poria S, Cambria E, Hazarika D, Mazumder N, Zadeh A, Morency L-P (2017) Context-dependent sentiment analysis in user-generated. ACL 2:873–883
7. Cambria E (2016) Affective computing and sentiment analysis. IEEE Intell Syst 31(2):102–107
8. Liu B, Zhang L (2012) A survey of opinion mining and sentiment analysis. In: Mining text data. Springer, Boston, pp 415–463
9. Huddar MG, Sannakki SS, Rajpurohit VS (2019) A survey of computational approaches and challenges in multimodal sentiment analysis. Int J Comput Sci Eng 7(1):876–883
10. Lo SL, Cambria E, Chiong R, Cornforth D (2017) Multilingual sentiment analysis: from formal to informal and scarce resource languages. Artif Intell Rev 48(4):499–527
11. Penga H, Ma Y, Lib Y, Cambria E (2018) Learning multi-grained aspect target sequence for Chinese sentiment analysis. Knowl-Based Syst 148:167–176
12. Mohammad SM, Kiritchenko S, Zhu X (2013) Building the state-of-the-art in sentiment analysis of tweets. In: Second Joint conference on lexical and computational semantics, Canada
13. Thakora P, Sasi DS (2015) Ontology-based sentiment analysis process for social media content. Procedia Comput Sci 53:199–207
14. Kiritchenko S, Zhu X, Mohammad SM (2014) Sentiment analysis of short informal texts. J Artif Intell Res 50:723–762
15. Nalisnick ET, Baird HS (2013) Extracting sentiment networks from Shakespeare's plays. In: 12th international conference on document analysis and recognition, Washington, DC, USA
16. Peng B, Li J, Chen J, Han X, Xu R, Wong K-F (2015) Trending sentiment-topic detection on twitter. In: International conference on intelligent text processing and computational linguistics
17. Lyu K, Kim H (2016) Sentiment analysis using word polarity of social media. Wireless Pers Commun 89(3):941–958
18. Gupta P, Tiwari R, Robert N (2016) Sentiment analysis and text summarization of online reviews: a survey. In: International conference on communication and signal processing (ICCSP), Melmaruvathur, India
19. de Kok S, Punt L, van den Puttelaar R, Ranta K, Schouten K, Frasincar F (2018) Review-aggregated aspect-based sentiment analysis with ontology features. Progress Artif Intell 7(4):295–306

20. Korayem M, Crandall D, Abdul-Mageed M (2012) Subjectivity and sentiment analysis of Arabic: a survey. In: International conference on advanced machine learning technologies and applications, Springer, Berlin
21. Ramteke J, Shah S, Godhia D, Shaikh A (2016) Election result prediction using Twitter sentiment analysis. In: International conference on inventive computation technologies (ICICT), Coimbatore, India
22. Mars A, Gouider MS (2017) Big data analysis to features opinions extraction of customer. Procedia Comput Sci 112:906–916
23. Li X, Xie H, Chen L, Wang J, Deng X (2014) News impact on stock price return via sentiment analysis. Knowl-Based Syst 69:14–23
24. Nagamma P, Pruthvi HR, Nisha KK, Shwetha NH (2015) An improved sentiment analysis of online movie reviews based on clustering for box-office prediction. In: International conference on computing, communication and automation, Noida, India
25. Kirilenko AP, Stepchenkova SO, Kim H, Li X (2018) Automated sentiment analysis in tourism: comparison of approaches. J Travel Res 57(8):1012–1025
26. Gohil S, Vuik S, Darzi A (2018) Sentiment analysis of health care tweets: review of the methods used. JMIR Public Health Surveill 4(2):e43
27. Chen LS, Huang TS, Miyasato T, Nakatsu R (1998) Multimodal human emotion/expression recognition. In: Proceedings of the 3rd international conference on face and gesture recognition, Washington, DC, USA
28. Wöllmer M, Weninger F, Knaup T, Schuller B, Sun C, Sagae K, Morency L-P (2013) YouTube movie reviews: sentiment analysis in an audio–visual context. IEEE Intell Syst 28(3):46–53
29. Wu CH, Liang WB (2010) Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. IEEE Trans Affect Comput 2(1):10–21
30. Noroozi F, Marjanovic M, Njegus A, Escalera S, Anbarjafari G (2017) Audio–visual emotion recognition in video clips. IEEE Trans Affect Comput 10(1):60–75
31. Rozgić V, Ananthakrishnan S, Saleem S, Kumar R, Prasad R (2013) 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
32. Huddar MG, Sannakki SS, Rajpurohit VS (2018) An ensemble approach to utterance level multimodal sentiment analysis. In: 2018 international conference on computational techniques, electronics and mechanical systems (CTEMS), Belgaum, India
33. Eyben F, Wöllmer M, Graves A, Schuller B, Douglas-Cowie E, Cowie R (2010) On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues. J Multimodal User Interfaces 3(1–2):7–19
34. Poria S, Chaturvedi I, Cambria E, Hussain A (2016) Convolutional MKL based multimodal emotion recognition and sentiment analysis. In: IEEE 16th international conference on data mining (ICDM), Barcelona, Spain
35. Zadeh A, Chen M, Poria S, Cambria E, Morency L-P (2017) Tensor fusion network for multimodal sentiment analysis. In: Empirical methods in natural language processing, Copenhagen, Denmark
36. Zadeh A, Zellers R, Pincus E, Morency L-P (2016) Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages. J IEEE Intell Syst 31(6):82–88
37. Busso C, Bulut M, Lee C, Kazemzadeh A, Mower E, Kim S, Chang J, Lee S, Narayanan S (2008) IEMOCAP: interactive emotional dyadic motion capture database. J Lang Resources Eval 42(4):335–359
38. Eyben F, Wöllmer M, Schuller B (2013) Recent developments in open SMILE, the munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM international conference on Multimedia, Barcelona, Spain
39. Mariethoz J, Bengio S (2005) A unified framework for score normalization techniques applied to text-independent speaker verification. IEEE Signal Process Lett 12(7):532–535
40. Karpathy A, Toderici G, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. In: Proceedings of international computer vision and pattern recognition
41. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv:1301.3781
42. Teh YW, Hinton GE (2000) Rate-coded restricted Boltzmann machines for face recognition. In: Proceedings of the 13th international conference on neural information processing systems, Cambridge, MA, USA
43. Ji S, Xu W, Yang M, Yu K (2013) 3d convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231