Hindawi
Security and Communication Networks
Volume 2021, Article ID 6662337, 10 pages
https://doi.org/10.1155/2021/6662337

Research Article

Two-Level Multimodal Fusion for Sentiment Analysis in Public Security

Jianguo Sun,[1] **Hanqi Yin**,[1] **Ye Tian**,[1] **Junpeng Wu**,[1] **Linshan Shen**,[1] **and Lei Chen**[2]

[1]*College of Computer Science and Technology, Harbin Engineering University, Harbin, Heilongjiang 150001, China*
[2]*College of Engineering and Computing, Georgia Southern University, Statesboro, GA 30458, USA*

Correspondence should be addressed to Linshan Shen; shenlinshan@hrbeu.edu.cn

Received 28 December 2020; Accepted 21 May 2021; Published 4 June 2021

Academic Editor: David Megías

Large amounts of data are widely stored in cyberspace. Not only can they bring much convenience to people's lives and work, but they can also assist the work in the information security field, such as microexpression recognition and sentiment analysis in the criminal investigation. Thus, it is of great significance to recognize and analyze the sentiment information, which is usually described by different modalities. Due to the correlation among different modalities data, multimodal can provide more comprehensive and robust information than unimodal in data analysis tasks. The complementary information from different modalities can be obtained by multimodal fusion methods. These approaches can process multimodal data through fusion algorithms and ensure the accuracy of the information used for subsequent classification or prediction tasks. In this study, a two-level multimodal fusion (TlMF) method with both data-level and decision-level fusion is proposed to achieve the sentiment analysis task. In the data-level fusion stage, a tensor fusion network is utilized to obtain the text-audio and text-video embeddings by fusing the text with audio and video features, respectively. During the decision-level fusion stage, the soft fusion method is adopted to fuse the classification or prediction results of the upstream classifiers, so that the final classification or prediction results can be as accurate as possible. The proposed method is tested on the CMU-MOSI, CMU-MOSEI, and IEMOCAP datasets, and the empirical results and ablation studies confirm the effectiveness of TlMF in capturing useful information from all the test modalities.

1. Introduction

The main way of human communication includes language, visual, and acoustic, which can be generally rendered by speaking or writing, images or videos, and encoded phoneme information, respectively. Therefore, it is incomplete and limited to study the opinion, attitude, and sentiment of an entity only through the language or vision modal. Focusing on this point, multimodal research has been popularized and widely used in many fields of machine learning and artificial intelligence, such as tasks of sentiment analysis [1, 2], emotion recognition [3, 4], behavior recognition [5], video captioning [6], and medical diagnose [7–9]. Besides, in the field of public information security, microexpression recognition [10, 11] is the core process in the lie detection task. However, since the microexpressions are short duration and hard to capture by human and machine, most existing investigations are unsatisfactory. Accordingly, using the information complementarity strategy, i.e., applying the voices and audio modalities of the video as auxiliary information in a microexpression recognition task, can effectively increase the accuracy of lie detection.

Multimodal fusion is the concept that integrates information of multiple modalities by classification or prediction [12] and has become one of the most popular research interests in the field of multimodal machine learning. The main advantages of multimodal fusion can be summarized as follows. First, it can obtain more accurate predictions by observing the phenomenon from different modalities. Second, a multimodal system can catch the information that is invisible in individual modalities by multimodal fusion. Third, a multimodal system can still operate when one of the

modalities is missing. In previous research studies on multimodal fusion, the early, late, and hybrid fusion methods [13–15] have been widely employed to achieve fusion tasks. Take the early fusion method as an example, which can achieve the fusion objective by connecting the embeddings of text (T), audio (A), and video (V). However, this method cannot effectively extract the interactive features among the modalities.

On the other hand, the neural network has been extensively used for the multimodal fusion tasks due to its impressive advantages. First, neural network approaches are adapted in learning from a large amount of data. Second, the existing architectures of the neural network allow the end-to-end training that is based on the multimodal representation component and the fusion component. Moreover, the neural network can find the optimal solution quickly, learn complex decision boundaries, and has the ability of fault tolerance. By means of neural networks, Zadeh et al. [16] proposed the tensor fusion network (TFN), which learns both the intra-modality and intermodality dynamics end-to-end. Compared with the early fusion method used for the concatenation of multiple modalities embeddings, the TFN can effectively learn interactions between different features by outer product. Based on the TFN, low-rank multimodal fusion (LMF) [17], memory fusion network (MFN) [18], and multimodal transformer (MuLT) [19] have been proposed, which can further improve the processing efficiency and evaluation. In addition, it can be seen from these results that attaching both audio and video features to the same textual information can enable nontext information to be better understood, and in turn, the nontext information can impart greater meaning to the text [20]. Therefore, the multimodal fusion can be achieved according to correlations between those text-based features.

It should be mentioned that the long short-term memory (LSTM) network [21] was utilized in [18, 22] to process the sequence of data. The basic component of the architecture is a simple LSTM cell that contains three gates, namely, input gate, forgot gate, and output gate. The input gate determines the information to be remembered, and the forget gate selects the information to be dropped. The function of the output gate determines the output from the input listed in the memory. Based on these gates, the LSTM network can provide a prediction at each time unit, and it can learn long-term information more easily than the traditional recurrent neural network (RNN) by resorting to its self-loop mechanism. It should be pointed out that the LSTM network can only utilize the past input features of a specific period. For the future one, which plays a key role in the time-dependent sentiment analysis task, it is unavailable. For this reason, the BiLSTM network [23] has been popularized since it can learn the backward and forward long-term dependencies between small timestamps of the data sequence simultaneously. For example, in the field of natural language processing (NLP), the BiLSTM network was utilized for the context representation task [24] where more accurate evaluation can be obtained than the LSTM-based results [25]. In [26], the authors completed the acoustic modeling task utilizing the BiLSTM network and showed the state-of-the-art results on the TIMIT speech database.

Fusion in the decision level is a higher level of information fusion, which combines information from different data sources classified individually [27]. Generally, decision fusion for multisources data has the potential to improve classification accuracy, compared with the individual data sources. In previous research studies, decision fusion is widely used in several classification tasks [28]. For example, in the remote sensing image classification task, Mazher et al. [29] proposed a decision fusion method in land cover classification. The authors indicate that the proposed decision fusion method has desired generalization performance and can be also applied to other applications with different sensor data. In the medical diagnosis task, Agarwal and Bedi [30] proposed a fusion method using curvelet and wavelet transform and applied it to the medical diagnosis task that combines the images obtained by computed tomography (CT) scan and magnetic resonance imaging (MRI). The experiment shows that the results of two-level fusion are better than those of one feature-level fusion and decision-level fusion. In the natural resource prediction task, DSS was applied to the problem of sustainable greenhouse management [31] and regarded as an effective strategy to balance the sustainability and the profitability of productions.

Motivated by the above discussion, this study proposes a new modal fusion method named TlMF, which produces unimodal embeddings by using a CNN-BiLSTM neural network and achieves the information fusion of text and audio/video embedding based on tensor fusion and decision fusion stages. The cores of our method include (1) a tensor fusion network used to fuse text data with video and audio data, respectively, and (2) a decision-level fusion strategy, which can fuse the classification results. Then, three datasets of multimodal sentiment analysis and emotion classification, i.e., CMU-MOSI [32], CMU-MOSEI [33], and IEMOCAP [34] are used for experiments to evaluate the effectiveness of the proposed methods.

The rest of this study is organized as follows: Section 2 introduces the proposed model; the experimental setup and the experimental results are given in Section 3, and the conclusion is provided in Section 4.

2. Proposed Method

Figure 1 shows the main diagram of the TlMF model that is established on the stages of data preparation, feature extraction, tensor fusion, and decision fusion. To make this framework clearer, detailed descriptions of some components are given in this section.

2.1. CNN-BiLSTM. A convolution neural network (CNN) is utilized to learn features from each modality after the data preparation stage. In this study, the input sequences of the text, audio, and video feature embeddings are defined as $H_t \in R^{d_t \times l_t}$, $H_a \in R^{d_a \times l_a}$, and $H_v \in R^{d_v \times l_v}$, respectively. To extract the features from these input sequences, a 1D temporary convolution layer is used for the time dimension of each input vector. And the outputs of this convolution layer are denoted as $H_{t1} \in R^{d_{t1} \times l_t}$, $H_{a1} \in R^{d_{a1} \times l_a}$, and

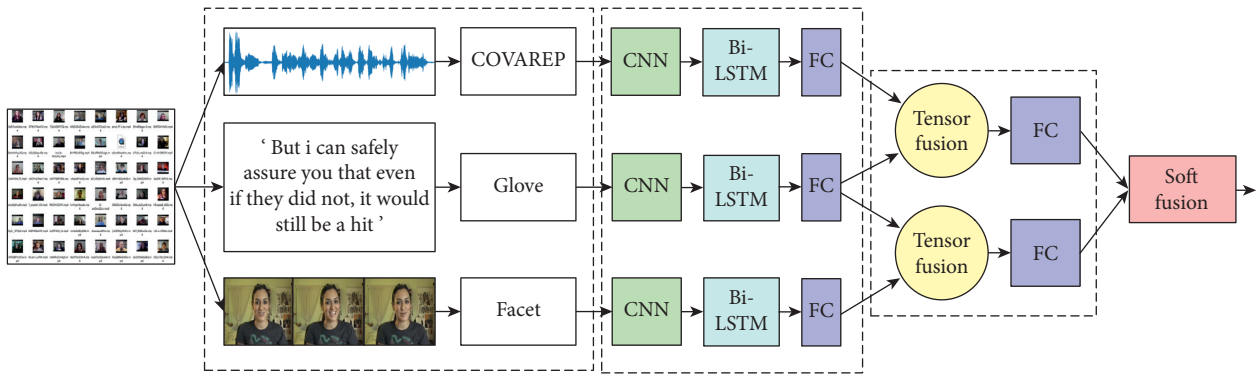

Figure 1: Overall architecture of TlMF. The first stage is data preparation, which turns the raw data into a unimodal sequence for text, audio, and video modalities. Once the unimodal sequence is obtained, the unimodal features can be learned by the second stage, which can extract features from each modality. Then, the tensor fusion layer is used to fuse the text-based audio feature H_{ta} and the text-based video feature H_{tv}. Finally, a decision fusion layer is employed to improve the accuracy of classification and prediction in the sentiment analysis task.

$H_{v1} \in R^{d_{v1} \times l_v}$, respectively. For ease of calculation, we assume $d_{t1} = d_{a1} = d_{v1}$.

The output of the CNN is then processed by the BiLSTM layer, which is extended from LSTM and can learn the context correlation from time-series data. It should be pointed out that LSTM has two special structures: cell state and gate. The latter consists of the input gate, the forget gate, and the output gate, which control the update, maintenance, and deletion of cell state, respectively. By denoting C_t as the current cell state of LSTM and h_t as the corresponding hidden state at the instant t, the forward propagation process of LSTM can be summarized as follows.

The first step is to update the output f_t of the forget gate, which controls whether to forget the cell state of the upper layer with a certain probability. Note that the input of the forget gate includes the hidden state h_{t-1} of the previous sequence and the current sequence data $x_t \in \{H_{t1}, H_{a1}, H_{v1}\}$. Thus, the rule of the output is described by the following equation:

$$f_t = \sigma\left(\theta_f \left[h_{t-1}, x_t\right] + b_f\right), \tag{1}$$

where θ_f and b_f are the sets of weight matrices and biases vectors in the forget gate, respectively; σ represents the sigmoid activation function.

The second step is to update the information stored in the current cell state C_t. For this purpose, two parts need to be considered. First, by sigmoid activation function, the output i_t of the input gate is generated, which contains the values to be updated. Then, a tanh layer creates a vector of new candidate values a_t that could be added to the state, as shown in the following equations:

$$i_t = \sigma\left(\theta_i \left[h_{t-1}, x_t\right] + b_i\right), \tag{2}$$

$$a_t = \tanh\left(\theta_a \left[h_{t-1}, x_t\right] + b_a\right), \tag{3}$$

where θ_i and b_i mean the sets of weight matrices and biases vectors in the input gate, respectively. Note that once f_t, a_t, and i_t are determined, the current cell state C_t can be updated from the last cell state C_{t-1} according to

$$C_t = \left(C_{t-1} \otimes f_t \oplus i_t \otimes a_t\right), \tag{4}$$

where $\otimes$ and $\oplus$ mean the Hadamard product and the concatenation operator, respectively.

Now, we can determine the desired output based on the cell state C_t. Let θ_o and b_o be the sets of weight matrices and biases vectors in the output gate, respectively. The main procedure is composed of two steps shown in equations (5) and (6). First, the output gate is to select the parts to be output from C_t, and the corresponding output is given by o_t. Then, by using the Hadamard product between $\tanh(C_t)$ and o_t, the desired output can be obtained.

$$o_t = \sigma\left(\theta_o \left[h_{t-1}, x_t\right] + b_o\right), \tag{5}$$

$$h_t = o_t \otimes \tanh\left(C_t\right). \tag{6}$$

Finally, the prediction output of the current sequence is updated in accordance with

$$y_n = \varphi\left(\theta_y h_t + b_y\right), \tag{7}$$

where θ_y and b_y represent the sets of weight matrices and biases vectors in the predict layer, respectively; φ is the output activation function, e.g., Softmax in classification task or tanh in prediction task.

So far, the forward propagation of LSTM has been described in detail. It can be seen from equations (1)–(7) that the LSTM network can only utilize the previous input features of a specific period. This characteristic of LSTM greatly restricts its application in the sentiment analysis task. In order to simultaneously learn the previous and future context from sequence data, BiLSTM is considered, which combines two separate hidden LSTM layers with the opposite directions of the same input.

The architecture of the BiLSTM network is shown in Figure 2, where $\overrightarrow{h}_t = (\overrightarrow{h}_{t_0}, \overrightarrow{h}_{t_1}, \ldots, \overrightarrow{h}_{t_n})$, $\overleftarrow{h}_t = (\overleftarrow{h}_{t_0}, \overleftarrow{h}_{t_1}, \ldots, \overleftarrow{h}_{t_n})$, and $x_t = (x_{t_0}, x_{t_1}, \ldots, x_{t_n})$. The main procedure of BiLSTM can be described by

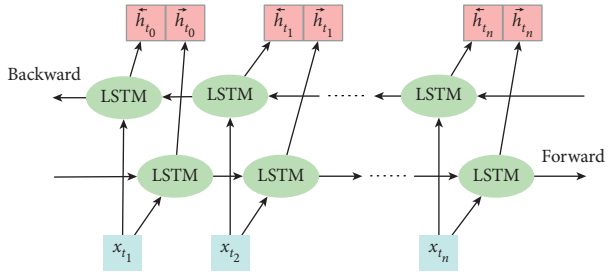

Figure 2: The architecture of the BiLSTM network.

$$\begin{aligned}
\overrightarrow{h}_t &= \sigma\left(\theta_{\overrightarrow{h}}\left[\overrightarrow{h}_{t-1}, x_t\right] + b_{\overrightarrow{h}}\right),\\
\overleftarrow{h}_t &= \sigma\left(\theta_{\overleftarrow{h}}\left[\overleftarrow{h}_{t-1}, x_t\right] + b_{\overleftarrow{h}}\right),\\
\left(\overrightarrow{h}_0, \overleftarrow{h}_0\right), \ldots, \left(\overrightarrow{h}_t, \overleftarrow{h}_t\right) &= \mathrm{BiLSTM}\left(x_0, x_1, \ldots, x_t\right),\\
y_t &= \sigma\left(\theta_{y_t \overrightarrow{h}}\overrightarrow{h}_t + \theta_{y_t \overleftarrow{h}}\overleftarrow{h}_t + b_{y_t}\right),
\end{aligned} \tag{8}$$

where $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ are the forward hidden sequence and the backward hidden sequence, respectively, and can be obtained from equations (5)–(7). $\theta_{\overrightarrow{h}}$, $\theta_{\overleftarrow{h}}$, $\theta_{y_t \overrightarrow{h}}\overrightarrow{h}_t$, and $\theta_{y_t \overleftarrow{h}}\overleftarrow{h}_t$ are the sets of weight matrices; $b_{\overrightarrow{h}}$, $b_{\overleftarrow{h}}$, and b_{y_t} represent the sets of biases vectors. The output vector $y_t = (y_0, y_1, \ldots, y_t)$ of the hidden layer is the concatenation of $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ and satisfies $y_t = [\overrightarrow{h}_t, \overleftarrow{h}_t]$.

In this study, each modal uses a BiLSTM block, which consists of three-stacked BiLSTM layers, two fully connection layers, a Maxpool layer, and a tanh layer. The output y_t from the upper layer becomes the input of the lower layer. Since the time series is limited, the computation load of the three BiLSTM layers will be not increased.

In the CNN-BiLSTM feature extraction network, the preprocessing is performed through a CNN subnetwork to learn local features, such that a shorter series with high-level features can be obtained. Then, a separate BiLSTM subnetwork is trained for different modalities. In this case, the output tensors of text, audio, and video are defined as $H_{t2} \in R^{d_{t2}}$, $H_{a2} \in R^{d_{a2}}$, and $H_{v2} \in R^{d_{v2}}$, respectively.

2.2. Tensor Fusion. In this study, the TFN [16] is adopted, which can effectively learn the interactions between different features by outer product. Moreover, the employment of outer product can lead to none learnable parameters and a low possibility of overfitting despite the high-dimensional output tensor. Therefore, we choose the outer product to fusion our data sequence.

It should be mentioned that since text modal contains more considerable sentiment-related information than video and audio modalities. For this reason, we fuse the text information with audio and video separately and denote text-audio and text-video features as H_{ta} and H_{tv} respectively. Specifically, after the CNN-BiLSTM network, we use the tensor fusion layer to learn interactions between different modalities. The inputs of tensor fusion layer are the outputs from the CNN-BiLSTM network. The calculation can be performed by the following equations:

$$H_{\mathrm{ta}} = \left[H_{t2} \otimes H_{a2}\right], \tag{9}$$

$$H_{\mathrm{tv}} = \left[H_{t2} \otimes H_{v2}\right]. \tag{10}$$

This process can be described by Figure 3. The fusion information $H_{\mathrm{ta}} \in R^{d_{t2} \times d_{a2}}$ and $H_{\mathrm{tv}} \in R^{d_{t2} \times d_{v2}}$ can be obtained after the tensor fusion layer. According to [16], since the output neurons of tensor fusion are easy to interpret and very meaningful in semantics, it is easy for the subsequent layers of the network to decode the meaningful information.

2.3. Decision Fusion. Fusion at the decision level between the classification results of the text-audio and text-video is considered to further improve the results. According to posterior probabilities, each classifier can yield a scoring matrix as the output of the Softmax layer to characterize the confidence level that the network chooses a specific class as the correct output class. Soft fusion is a method to get the new prediction label by fusing the scoring matrix from classifiers. Since soft fusion can increase accuracy in the minimum and worst case, it has been widely used in the medical diagnose field [9, 35]. In this study, the soft fusion method, which can incorporate the scores from separate fusion modalities, is applied to generate the new prediction label in the sentiment analysis task. The weighted combination of the two fusion modalities scores is shown in the following equation:

$$S_F(c) = W_{\mathrm{ta}} \cdot S_{\mathrm{ta}}(c) + W_{\mathrm{tv}} \cdot S_{\mathrm{tv}}(c), \tag{11}$$

where W_{ta} and W_{tv} denote the weight of text-audio modal and text-video modal, respectively, and satisfy $W_{\mathrm{ta}} = W_{\mathrm{tv}}$ at the initial time. $S_{\mathrm{ta}}(c)$ is the score matrix of the text-audio modal for the prediction of class c, $S_{\mathrm{tv}}(c)$ represents the score matrix of the text-video modal, and $S_F(c)$ stands for the final classification results.

3. Experiment and Discussion

3.1. Datasets. The proposed method is tested by using three public multimodal datasets: CMU Multimodal Corpus of Sentiment Intensity (CMU-MOSI) [32] and CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) [33] datasets for sentiment analysis task and Interactive Emotional Dyadic Motion Capture (IEMOCAP) [34] dataset for emotion classification task. The detailed descriptions of these datasets are given as follows.

CMU-MOSI is built on 93 YouTube movie review videos that are segmented into 3,702 utterance segments including 2,199 opinion video segments. Each segment is annotated by a scale in (−3, 3) to reflect sentiment intensity, where −3 and 3 represent the extremely negative and extremely positive sentiments, respectively. These segments are rigorously annotated

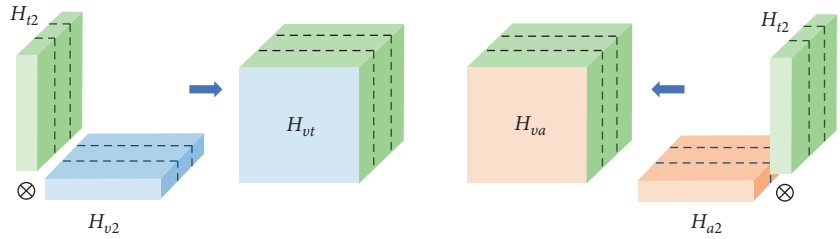

Figure 3: Tensor fusion.

according to the subjective sentiment intensity of the visual features and audio features. Moreover, these segments divided into three parts such as train, validation, and test are 16326, 1871, and 4659 data samples, respectively. The raw data and the distribution of sentiment intensity are shown in Figures 4 and 5, respectively. From Figure 5, we can clearly figure out that the number of segments in different sentiment intensity is relatively average.

CMU-MOSEI is the largest dataset of multimodal sentiment analysis tasks. It consists of 23453 sentence utterance video segments from more than 1000 online YouTube speakers and 250 topics. Each segment video is transcribed and properly punctuated, which can be treated as an individual multimodal example. In this case, train, validation, and test partitions contain 16326, 1871, and 4659 data samples, respectively. Figure 6 shows the part of raw data, and the sentiment intensity distributed in (−3, 3) is given in Figure 7, where −3 and 3 represent the extremely negative and extremely positive sentiments, respectively. It is clear from Figure 7 that most segments are labeled with (0, 1), which implies the sentiment intensity is weakly positive. The uneven distribution of the sentiment intensity corresponds to people's commenting habits.

IEMOCAP is an acted, multimodal, and multispeaker database, which has a free academic license, long duration, and good emotion label. It contains about 12 hours of 302 videos, in which speakers performed 9 different emotions: angry, excited, fear, sad, surprised, frustrated, happy, disappointed, and neutral. In this study, happy, sad, angry, and neutral emotions are chosen for experiment, and the distribution is shown in Figure 8. These considered videos are divided into 4444 segments with emotion annotations. In this case, train, validation, and test partitions contain 2717, 789, and 938 data samples, respectively.

3.2. Multimodal Features. For the above datasets, the unimodal features are extracted from the text, audio, and video data modalities by using global vectors for word representation (Glove) [36], COVAREP [37], and Facet, respectively. Moreover, these unimodal features can also be provided by the CMU-MultimodalSDK.

Text features is extracted by Glove, which is an unsupervised learning algorithm for obtaining vector representations of words. The dimension of each text embedding extracted by Glove is 300 for different inputs in the above datasets.

Audio features are extracted by utilizing COVAREP, which is a collaborative and freely available repository of speech processing algorithms. By using COVAREP, we can obtain low-level acoustic features including 12 Mel-frequency cepstral coefficients, pitch tracking, voiced/unvoiced segmenting features, glottal source parameters, and peak slope parameters. Each audio feature is extracted on 25-ms frames with a 5-ms shift, and its dimension is 74 for each dataset.

Video features consist of 35 facial action units extracted from each frame by using Facet, which is widely used for extracting facial expression features such as basic and advanced emotions. Thus, the dimension of each video frame feature is 35 for each dataset.

3.3. Performance Metrics. In this section, to show the advantages of the proposed method over the existing methods [16, 17, 19] in multiclass classification and prediction tasks, the following performance metrics are chosen for different datasets.

The CMU-MOSI and CMU-MOSEI datasets are labeled in the range of (−3, 3). According to the suggestion of the authors of the datasets, these labels can be divided into different groups. In this paper, two groups are considered. The labels in the first group are divided into two categories: the negative sentiment and the positive sentiment, which are with the labels in the range of (−3, 0) and (0, 3), respectively. In this case, we choose the binary accuracy (Acc-2) and F-score as the performance metrics on the CMU-MOSI and CMU-MOSEI datasets. The 7-class accuracy (Acc-7) is used as the performance metrics in the other group which have 7 sentiment score classification in (−3, 3). In addition, the mean absolute error (MAE) and the correlation between the prediction results and the ground truth label are chosen in these two datasets.

The IEMOCAP dataset are labeled in 9 different emotions, where happy, sad, angry, and neutral emotions are chosen in this study. For this emotion classification task, F-score and the binary accuracy (Acc-2) are selected as the performance metrics in each emotion.

3.4. Training Setup. The proposed method is implemented by open-source PyTorch framework, and it is tested and evaluated on the computer with Intel (R) Xeon (R) Silver 4214 at 2.20 GHz CPU, TITAN RTX GPU with 24 GB memory, and 64 GB computer memory. In addition, the hyperparameter is configured for different datasets, where the learning rate and the batch size vary in the range of

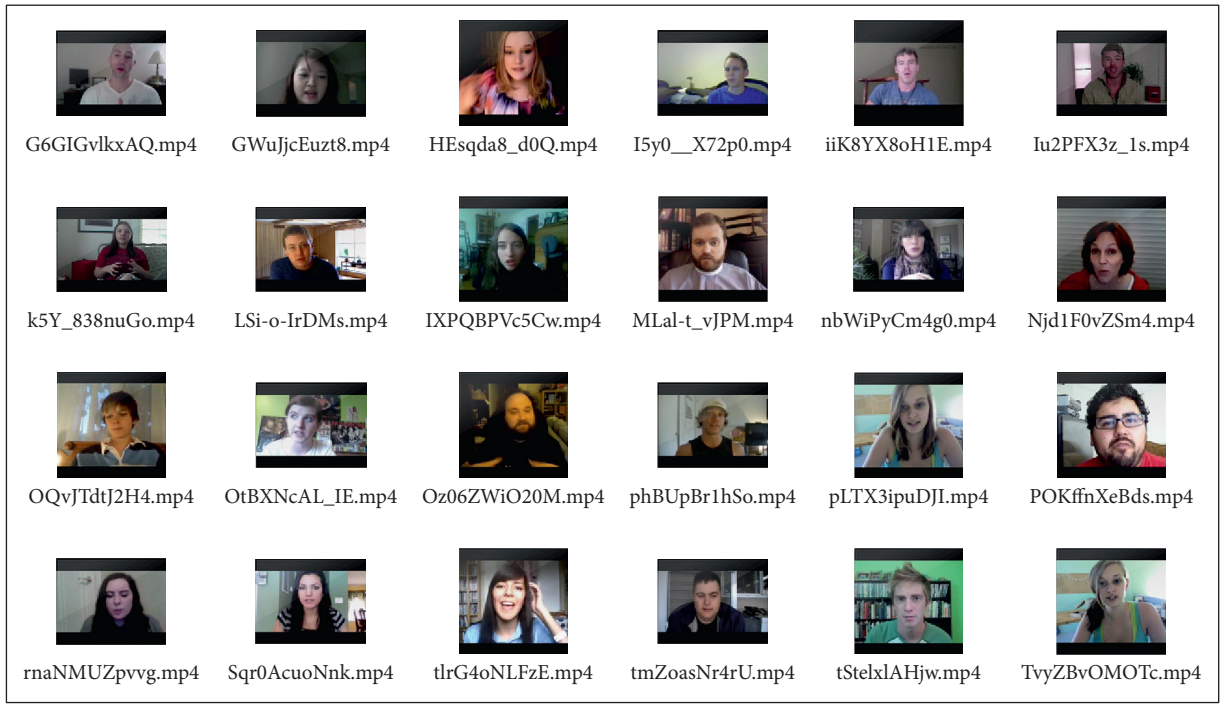

Figure 4: Part of raw videos in the CMU-MOSI dataset.

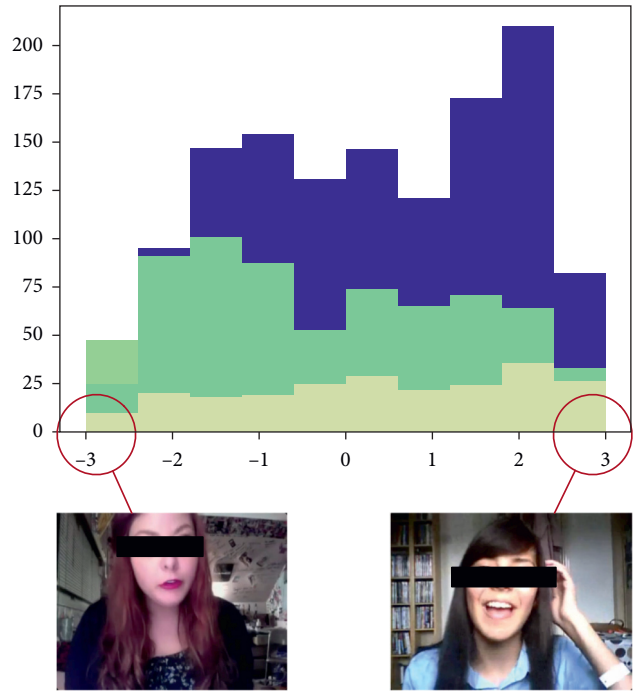

Figure 5: The sentiment intensity distribution of the CMU-MOSI dataset. Blue, yellow, and green represent train, valid, and test data. −3 and 3 represent the extremely negative and extremely positive sentiments, respectively.

$(10^{-3}, 10^{-5})$ and $(16, 128)$, respectively, and the epochs are trained 40–100 times by the model. ReLU and sigmoid are used as the activation functions. Mean square error is used as the loss function, and Adam is used as the optimizer. After optimizing the loss function, we use back propagation to update the parameters in the whole network.

3.5. Comparison between TlMF and the Existing Methods. In this subsection, the comparison between the proposed method and the existing methods, i.e., the tensor fusion network (TFN) [16], low-rank multimodal fusion (LMF) [17], and multimodal transformer (MuLT) [19], is made based on the experimental results of the sentiment analysis

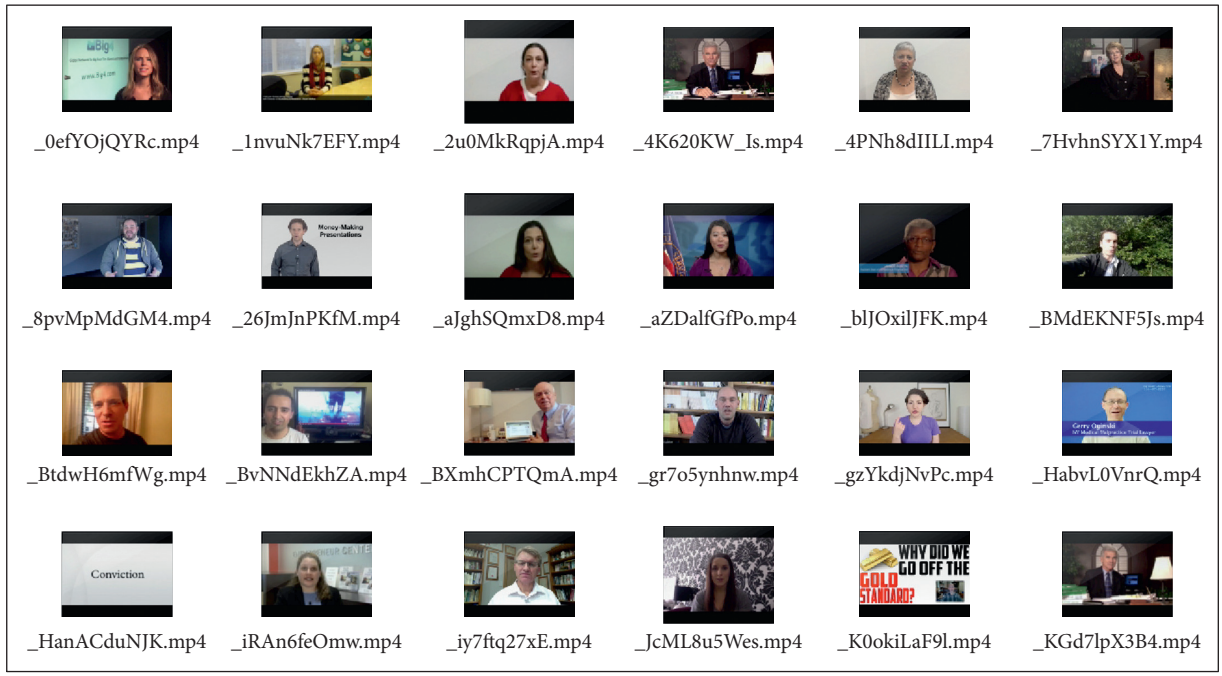

Figure 6: Part of raw videos in the CMU-MOSEI dataset.

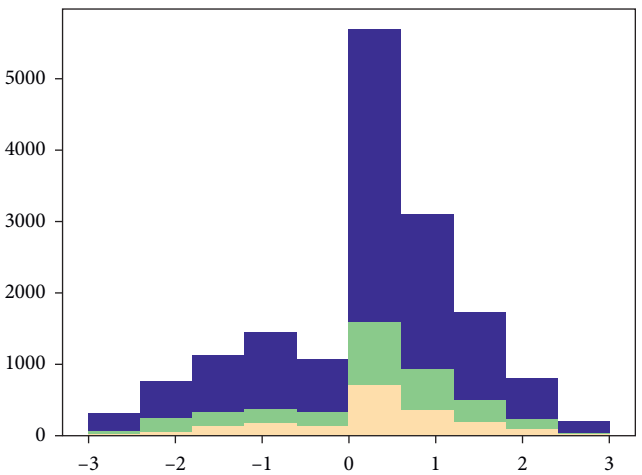

Figure 7: The sentiment intensity distribution of the CMU-MOSEI dataset. Blue, yellow, and green represent train, valid, and test data, respectively.

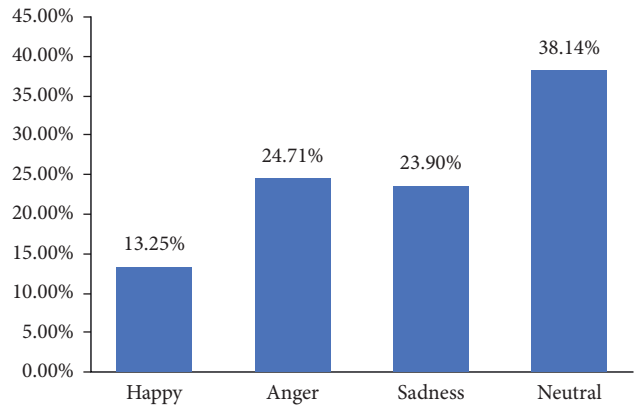

Figure 8: Emotion distribution of the IEMOCAP dataset.

task. It is known from [16] that the TFN combines individual modal's embeddings via calculating three different outer product subtensors such as unimodal, bimodal, and trimodal, and all subtensors are to be flattened as multimodal embedding vectors. As for LMF, it learns the multimodal embedding based on the similar tensor processing to that of the TFN, but with an additional low-rank factor for reducing computation memory. Different from the TFN and LMF, MuLT focus on interactions between multimodal sequences and latently adapt streams from one modality to another and uses the currently popular transformer model to transform one modal into another for modal fusion. In [19], MuLT has achieved state-of-art results in recent years on the multimodal fusion field. Thus, the comparison with MuLT can further test the superiority of TlMF on the sentiment analysis and the emotion classification. To guarantee the comparison fair, the three methods are tested under the same features and experimental environment.

In teams of the sentiment analysis task, the experiments are performed based on the CMU-MOSI and CMU-MOSEI datasets, and the corresponding comparison results are given in Table 1, respectively. From Table 1, it is seen that TlMF can obtain higher Acc-2, F-score, Acc-7, and Corr and lower MAE than TNF, LMF, and MuLT in the CMU-MOSI dataset. In the CMU-MOSEI dataset, it is also seen that TlMF can obtain higher Acc-2, F-score, Acc-7, and Corr than TNF, LMF, and MuLT. In other words, the proposed method is more effective in achieving accurate prediction than the existing methods under these two datasets. On the other hand, the comparison experimental results on the IEMOCAP datasets are given in Table 2. By observing this table, it is clear that the proposed method leads to higher

Table 1: Comparison experimental results for sentiment analysis on the CMU-MOSI and CMU-MOSEI datasets.

Metric	Acc-2	F-score	MAE	Acc-7	Corr
Case of the CMU-MOSI dataset					
TFN	80.82	80.77	0.901	23.94	0.698
LMF	82.53	82.47	0.917	33.23	0.695
MuLT	79.11	79.10	0.977	33.23	0.666
TlMF	92.28	92.26	0.373	65.49	0.933
Case of the CMU-MOSI dataset					
TFN	78.57	78.09	0.623	47.21	0.648
LMF	78.03	78.18	0.609	48.02	0.657
MuLT	79.35	79.29	0.631	48.45	0.647
TlMF	79.46	79.46	0.645	48.88	0.669

Table 2: Comparison experimental results for emotion classification on IEMOCAP.

Emotions	Happy		Sad		Angry		Neutral	
Metric	ACC-2	F-score	ACC-2	F-score	ACC-2	F-score	ACC-2	F-score
TFN	82.66	82.03	80.63	80.75	82.11	82.03	65.03	65.90
LMF	82.62	83.38	79.53	80.15	82.62	83.38	63.48	67.05
MuLT	85.60	78.96	79.42	70.31	75.79	65.36	59.59	50.33
TlMF	97.86	97.81	96.98	96.96	97.82	97.81	96.68	96.65

Acc-2 and F-score than the existing method in different emotions, which implies that more accurate classification can be achieved by TlMF. From the above discussion, it can be concluded that the proposed method has great superiority on accuracy of prediction and classification over the TFN, LMF methods when it is used for sentiment analysis.

3.6. Ablation Study. In this subsection, two series of ablation studies that compare different TlMF's variants are performed on the basis of the CMU-MOSI, CMU-MOSEI, and IEMOCAP datasets. The first series is to verify the effectiveness of the data fusion strategy in tensor fusion. For this purpose, we set text, audio, and video as the method's input separately and neglect the tensor fusion and decision fusion, such that the outputs corresponding to different inputs can be directly obtained from the downstream classifier. In this case, the text-only, audio-only, and video-only variants are denoted by TlMF_t, TlMF_a, and TlMF_v, respectively. In the second series of ablation studies, the influence of the tensor fusion and the decision fusion on the performance of the proposed method is illustrated by the following two cases:

Case 1: set text, audio, and video as the input. Discard the tensor fusion from TlMF and denote this variant as TlMF_{ntf}. The effectiveness of the tensor fusion can be confirmed by comparing the performance metrics of TlMF_{ntf} and TlMF.

Case 2: set the input as the same as in Case 1 and discard the decision fusion from TlMF. To keep the dimension of the output unchanged, the average-fusion is adopted to fuse the classification results of the upstream classifiers. In this case, the variant is defined as TlMF_{avg}, and the effectiveness of the decision fusion can be verified by comparing the performance metrics of TlMF_{avg} and TlMF.

Based on the above discussion, the experimental results of the ablation studies are provided in Tables 3 and 4. Specifically, Table 3 depicts the comparison results between TlMF and its variants in the sentiment analysis task when different datasets are utilized. It can be seen from the first three rows of these two subtables that Acc-2 and F-score of the text-only method are larger than those of the audio-only and video-only methods, which implies that the text modal can achieve more accurate sentiment analysis than the audio modal and video modal. The main contribution to this phenomenon is that the intrinsic structure of text is more adapted to emotional expression. Thus, it can represent language information better than the other modalities in the sentiment analysis task. On the other hand, by comparing the results of TlMF_{ntf} and TlMF (Case 1), it is clear that with the utilization of the tensor fusion, Acc-2, F-score, Acc-7, and MAE are, respectively, improved by 2.6%, 0.44%, 4.2%, and 4.2% on the CMU-MOSI dataset and by 3.96%, 4.1%, 5.01%, and 8.8% on the CMU-MOSEI dataset, which fully demonstrates the advantage of the tensor fusion in improving prediction accuracy of the sentiment analysis. Moreover, following a similar line to the analysis of Case 1, the effectiveness of the decision fusion can be confirmed according to the comparison between TlMF_{avg} and TlMF (Case 2).

For the emotion classification task, Table 4 provides the comparison results between TlMF and its variants on different emotions. By observing this table, there are two points worth emphasizing. First, it is seen from the first three rows that the text-only method leads to higher Acc-2 and F-score than the other methods for different emotions, that is, the text modal can achieve more accurate emotion classification than the audio modal and video modal. Second, by comparing the performance metrics of TlMF_{ntf} in Case 1 and TlMF_{avg} in Case 2 with those of TlMF, the effectiveness of the tensor fusion and the decision fusion can be clearly verified,

Table 3: Ablation experimental results for sentiment analysis on the CMU-MOSI and CMU-MOSEI datasets.

Metric	Acc-2	F-score	MAE	Acc-7	Corr
Case of the CMU-MOSI dataset					
TlMF_t	89.43	89.41	0.534	56.15	0.869
TlMF_a	77.33	77.31	0.983	30.37	0.51
TlMF_v	79.52	79.86	0.810	41.35	0.683
TlMF_{ntf}	90.08	90.06	0.512	56.15	0.886
TlMF_{avg}	89.60	89.62	0.553	52.95	0.886
TlMF	92.28	92.26	0.373	65.49	0.933
Case of the CMU-MOSI dataset					
TlMF_t	78.59	78.74	0.647	47.95	0.611
TlMF_a	62.27	64.52	0.717	41.36	0.593
TlMF_v	64.85	68.59	0.706	42.58	0.537
TlMF_{tnf}	75.53	76.86	0.716	44.53	0.534
TlMF_{avg}	77.35	77.97	0.665	46.98	0.581
TlMF	79.46	79.46	0.645	48.88	0.669

Table 4: Ablation experimental results for emotion classification on the IEMOCAP dataset.

Emotions	Happy		Sad		Angry		Neutral	
Metric	Acc-2	F-score	Acc-2	F-score	Acc-2	F-score	Acc-2	F-score
TlMF_t	97.46	97.41	95.14	95.08	94.58	94.54	93.48	93.42
TlMF_a	87.55	81.75	74.60	63.75	77.95	75.26	67.90	62.87
TlMF_v	94.11	93.40	92.08	91.62	91.97	91.58	86.52	85.84
TlMF_{ntf}	96.61	96.57	87.67	87.55	90.79	90.76	91.24	90.97
TlMF_{avg}	97.16	97.03	95.73	95.68	97.05	97.03	95.98	95.97
TlMF	97.86	97.81	96.98	96.96	97.82	97.81	96.68	96.65

respectively, due to the improvement of Acc-2 and F-score in different emotions. Thus, it can be concluded that the use of these two fusion strategies can effectively improve the accuracy of the emotion classification.

4. Conclusion

Harmful information such as extreme emotions and potential violence widely exists in multimodal data that are stored in cyberspace. Its analysis are great necessary and urgent in the field of the public information security. To this end, this study has proposed a novel TlMF method that contains both data-level and decision-level fusion. A CNN-BiLSTM neural network has been employed to produce the unimodal embeddings. Based on this, the text-audio and text-video fusion embeddings can be then generated by a tensor fusion network. Furthermore, the soft fusion method has been introduced in the decision-fusion stage of TlMF, such that the classification or prediction results can be as accurate as possible. Finally, the testing results on multimodal sentiment analysis and emotion classification tasks have confirmed that the multimodal features learned from the TlMF model can lead to state-of-the-art performance. This fully shows the effectiveness of the proposed method. Moreover, the availability of the network's components has been verified through the ablation studies.

Data Availability

The data used to support the findings of this study are found at http://immortal.multicomp.cs.cmu.edu/raw_datasets/processed data/.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by Fundamental Research Funds for the Central Universities (3072020CFQ0602, 3072020CF0604, and 3072020CFP0601) and in part by 2019 Industrial Internet Innovation and Development Engineering (KY1060020002 and KY10600200008).

References

[1] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: harvesting opinions from the web," in *Proceedings of the 13th International Conference on Multimodal Interfaces*, pp. 169–176, Alicante, Spain, November 2011.

[2] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image and Vision Computing*, vol. 65, pp. 3–14, 2017.

[3] W. Wang, *Machine Audition: Principles, Algorithms, and Systems*, IGI Global, Hershey, PA, USA, 2011.

[4] Y. Huang, J. Yang, P. Liao, and J. Pan, "Fusion of facial expressions and EEG for multimodal emotion recognition," *Computational Intelligence and Neuroscience*, vol. 2017, Article ID 2107451, 8 pages, 2017.

[5] D. Wu, J. Chen, W. Deng, Y. Wei, H. Luo, and Y. Wei, "The recognition of teacher behavior based on multimodal information fusion," *Mathematical Problems in Engineering*, vol. 2020, Article ID 8269683, 8 pages, 2020.

[6] S. Lee and I. Kim, "Multimodal feature learning for video captioning," *Mathematical Problems in Engineering*, vol. 2018, Article ID 3125879, 8 pages, 2018.

[7] B. Huang, F. Yang, M. Yin, X. Mo, and C. Zhong, "A review of multimodal medical image fusion techniques," *Computational and Mathematical Methods in Medicine*, vol. 2020, p. 16, 2020.

[8] S. El-Sappagh, T. Abuhmed, S. M. Riazul Islam, and K. S. Kwak, "Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data," *Neurocomputing*, vol. 412, pp. 197–215, 2020.

[9] H. Li, A. Shrestha, H. Heidari, J. Le Kernec, and F. Fioranelli, "Bi-LSTM network for multimodal continuous human activity recognition and fall detection," *IEEE Sensors Journal*, vol. 20, no. 3, pp. 1191–1201, 2019.

[10] S.-J. Wang, H.-L. Chen, W.-J. Yan, Y.-H. Chen, and X. Fu, "Face recognition and micro-expression recognition based on discriminant tensor subspace analysis plus extreme learning machine," *Neural Processing Letters*, vol. 39, no. 1, pp. 25–43, 2014.

[11] D. Patel, X. Hong, and G. Zhao, "Selective deep features for micro-expression recognition," in *Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR)*, Cancun, Mexico, December 2016.

[12] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: a survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[13] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Computing Surveys (CSUR)*, vol. 47, no. 3, pp. 1–36, 2015.

[14] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2008.

[15] C. G. Snoek, M. Worring, and A. W. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th Annual ACM International Conference on Multimedia*, pp. 399–402, Singapore, November 2005.

[16] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," 2017, https://arxiv.org/abs/1707.07250.

[17] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proceedings of the 56th Annual Meeting of the Association-For-Computational-Linguistics (ACL)*, Melbourne, Australia, July 2018.

[18] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, New York, NY, USA, February 2018.

[19] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the Conference Association for Computational Linguistics Meeting*, p. 6558, Florence, Italy, July 2019.

[20] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, pp. 8992–8999, 2020.

[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[22] A. Zadeh, C. Mao, K. Shi et al., "Factorized multimodal transformer for multimodal sequential learning," 2019, https://arxiv.org/pdf/1911.09826.pdf.

[23] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks: The Official Journal of the International Neural Network Society*, vol. 18, no. 5-6, pp. 602–610, 2005.

[24] O. Melamud, J. Goldberger, and I. Dagan, "context2vec: learning generic context embedding with bidirectional LSTM," in *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 51–61, Berlin, Germany, August 2016.

[25] A. A. Sharfuddin, M. N. Tihami, and M. S. Islam, "A deep recurrent neural network with BiLSTM model for sentiment classification," in *Proceedings of the 2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pp. 1–4, Sylhet, Bangladesh, September 2018.

[26] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 273–278, Olomouc, Czech Republic, December 2013.

[27] J. A. Benediktsson and I. Kanellopoulos, "Classification of multisource and hyperspectral data based on decision fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 3, pp. 1367–1377, 1999.

[28] B. Waske and J. A. Benediktsson, *Decision Fusion, Classification of Multisource Data*, Springer, Berlin, Germany, 2014.

[29] A. Mazher, P. Li, T. A. Moughal, and H. Xu, "A decision fusion method using an algorithm for fusion of correlated probabilities," *International Journal of Remote Sensing*, vol. 37, no. 1, pp. 14–25, 2016.

[30] J. Agarwal and S. S. Bedi, "Implementation of hybrid image fusion technique for feature enhancement in medical diagnosis," *Human-centric Computing and Information Sciences*, vol. 5, no. 1, pp. 1–17, 2015.

[31] G. Aiello, I. Giovino, M. Vallone, P. Catania, and A. Argento, "A decision support system based on multisensor data fusion for sustainable greenhouse management," *Journal of Cleaner Production*, vol. 172, pp. 4057–4065, 2018.

[32] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," 2016, https://arxiv.org/abs/1606.06259.

[33] 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*, pp. 2236–2246, Melbourne, Australia, July 2018.

[34] C. Busso, M. Bulut, C.-C. Lee et al., "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[35] H. Li, A. Shrestha, H. Heidari, J. L. Kernec, and F. Fioranelli, "A multisensory approach for remote health monitoring of older people," *IEEE Journal of Electromagnetics, RF and Microwaves in Medicine and Biology*, vol. 2, no. 2, pp. 102–108, 2018.

[36] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Jeju, South Korea, March 2014.

[37] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 960–964, Florence, Italy, May 2014.