Intelligent Automation & Soft Computing

Tech Science Press

DOI: 10.32604/iasc.2023.038835
Article

Multimodal Sentiment Analysis Using BiGRU and Attention-Based Hybrid Fusion Strategy

Zhizhong Liu*, Bin Zhou, Lingqiang Meng and Guangyu Huang

School of Computer and Control Engineering, Yantai University, Yantai, 264005, China
*Corresponding Author: Zhizhong Liu. Email: zhizhongLiu@ytu.edu.cn

Received: 30 December 2022; Accepted: 12 April 2023; Published: 23 June 2023

Abstract: Recently, multimodal sentiment analysis has increasingly attracted attention with the popularity of complementary data streams, which has great potential to surpass unimodal sentiment analysis. One challenge of multimodal sentiment analysis is how to design an efficient multimodal feature fusion strategy. Unfortunately, existing work always considers feature-level fusion or decision-level fusion, and few research works focus on hybrid fusion strategies that contain feature-level fusion and decision-level fusion. To improve the performance of multimodal sentiment analysis, we present a novel multimodal sentiment analysis model using BiGRU and attention-based hybrid fusion strategy (BAHFS). Firstly, we apply BiGRU to learn the unimodal features of text, audio and video. Then we fuse the unimodal features into bimodal features using the bimodal attention fusion module. Next, BAHFS feeds the unimodal features and bimodal features into the trimodal attention fusion module and the trimodal concatenation fusion module simultaneously to get two sets of trimodal features. Finally, BAHFS makes a classification with the two sets of trimodal features respectively and gets the final analysis results with decision-level fusion. Based on the CMU-MOSI and CMU-MOSEI datasets, extensive experiments have been carried out to verify BAHFS's superiority.

Keywords: Multimdoal sentiment analysis; BiGRU; attention mechanism; features-level fusion; hybrid fusion strategy

1 Introduction

With the development and maturity of social media, people like to express their opinions by posting videos, images and audio on social media platforms. The form of social media data is no longer limited to text modality but presents multiple data modalities. In recent years, sentiment analysis methods based on unimodal data have achieved good achievements in user satisfaction analysis, public opinion monitoring and other aspects. However, these methods cannot effectively deal with multimodal data and cannot make full use of diversity information hidden in multimodal data. To address this issue, multimodal sentiment analysis has been proposed and becomes a new research hotspot [1], which aims to mine users' opinions and sentiment states from multimodal data such as text, audio or visual based on unimodal sentiment analysis [2]. Actually, multimodal data [3] implies massive sentiment information, and it is useful for improving the accuracy of sentiment analysis by effectively integrating and learning the multimodal data.

In addition to capturing the internal features of unimodal data, multimodal sentiment analysis pays more attention to how to integrate the features of unimodal data (e.g., text, audio and video) to obtain a more precise representation of multimodal data, thus elevating the effectiveness of the multimodal sentiment analysis model. Now, there are two common fusion mechanisms for multimodal data fusion, which are the feature-level fusion mechanism [4,5] and the decision-level fusion mechanism [6,7]. The feature-level fusion mechanism usually concatenates the feature vectors of multiple modalities [8] to form a multimodal feature vector, thus implementing the interaction of multimodal data. Peng et al. [4] proposed a cross-modal complementary network applying the hierarchical method to fuse representations within and between modalities. Deng et al. [9] applied the long short term memory (LSTM) model to draw visual information over time, which is then fused with audio and textual cues to identify sentiment. Zadeh et al. [10], design a memory fusion network and used it to realize multi-view modeling with the concatenate method, however, this method leads to overfitting in a small-scale training dataset and ignores the specific dynamics of the modeling view. Ghosal et al. [11] proposed a bimodal attention fusion method, which fuses any two modality features with attention to capture the interaction information between the dual modalities. However, this model cannot capture the interaction information between the three modalities at the same time. When concatenating the feature vectors belonging to different modality data to form a high-dimensional feature vector, the feature-level fusion may ignore the implicit correlation between different modalities [12]. In addition, existing feature-level fusions may not capture the correlations between all modal features simultaneously.

In contrast to feature-level fusion which focuses on using a single classifier to handle high-dimensional feature vectors, decision-level fusion prefers to use multiple classifiers for decision-making and combine multiple decisions into one ensemble decision. For decision-level fusion, each modality is first analyzed separately, and then their decisions are combined through averaging, majority voting [13] and weighted summation [14]. Wang et al. [15] proposed a selective additive learning (SAL) method to improve the generalization ability of the multimodal sentiment analysis model. For the selection phase, SAL identifies confounding factors from the latent representations. In the addition stage, SAL adds gaussian noise to these representations so that the original model can discard confounding elements. Yu et al. [16] proposed a sentiment analysis method for microblog vision and text based on a deep convolution neural network (CNN). In this method, CNN and deep neural network (DNN) are used to analyze the sentiment of text information and visual information, respectively. Finally, the average strategy and weighting strategy are used to fuse the analysis results of the two modalities at the decision-level. The decision-level fusion mechanism has the advantages of a high degree of modularity, it can construct a multimodal model from a pre-trained unimodal model by fine-tuning the output layer, and always has better performance than unimodal models. However, the decision-level fusion mechanism has comings in cross-modal modeling and cannot capture the relationship between multiple modalities well.

The hybrid fusion strategy is composed of feature-level fusion and decision-level fusion, which extends the advantages of feature-level fusion and decision-level fusion. Gunes et al. [17] proposed a visual-based multimodal emotion recognition framework, which first implements the feature-level fusion after recognizing the facial expressions and gesture features from visual sequences, and then conducts the decision-level fusion with the product and weighting methods. Nemati et al. [18] proposed a multimodal hybrid fusion method, which fuses the audio and visual modalities with the latent spatial linear mapping and then adopts the fusion method based on evidence theory to project it to the cross-modality fusion of spatial features and text modalities. However, existing hybrid fusion methods first perform a feature-level fusion of two modalities and then perform decision-level fusion with a single modality that is not included, which cannot capture all modality features during feature-level and decision-level fusion.

To tackle the above issues, we present a multimodal sentiment analysis model using BiGRU and attention-based hybrid fusion strategy (BAHFS). Firstly, BAHFS first applies bidirectional gated recurrent units (BiGRU) to learn the deep context features inside the unimodal and then feeds the unimodal features into the bimodal attention fusion module to obtain the fused bimodal features. After that, the fused bimodal features and all the unimodal features are sent into the trimodal concatenation fusion module and the trimodal attention fusion module to obtain two sets of fused trimodal features, respectively. Subsequently, the two sets of trimodal features are classified separately and the analysis results are fused at the decision-level. Finally, the result of sentiment analysis is obtained after decision-level fusion. The contributions of this work are summarized as follows:

- To tackle the issue that existing feature-level fusion methods cannot capture the interaction information between the three modes simultaneously, we propose a trimodal attention fusion method that fuses bimodal features and unincluded unimodal features to capture the latent information between the three modalities. It is worth noting that the method is progressive, i.e., the output of the bimodal attention fusion and the unincluded unimodal features are used as the input of the trimodal attention fusion method.
- To tackle the issue that existing hybrid fusion methods do not capture the connections between all modal features during feature-level and decision-level fusion, we propose a new hybrid fusion strategy including feature-level fusion and decision-level fusion. For feature-level fusion, we adopt trimodal concatenation fusion and trimodal attention fusion methods to capture the connections between the trimodal features; For decision-level fusion, the maximum rule is used to fuse the two sets of trimodal features after feature-level fusion. This hybrid fusion strategy simultaneously captures the interaction information between all modalities during feature-level and decision-level fusion.
- We conduct an extensive experiment based on two public datasets CMU-MOSI and CMU-MOSEI, and the performance superiority of our proposed model is verified by comparing it with other 11 new and typical multimodal sentiment analysis models.

2 Related Work

Recently, with the rapid development of social media, users are now more inclined to express their opinions through multimodal data like text, audio, and video. In fact, multimodal data contain richer information, and the accuracy of sentiment analysis can be improved by learning from multimodal data. Currently, multimodal sentiment analysis has become a hot topic in the field of artificial intelligence. Traditional research about sentiment analysis mostly applies feature-based classification methods. Cao et al. [19] used an emotional dictionary to extract text features, uses adjective-noun pairs (ANP) to extract visual features in images, and then the text and image features are fused for sentiment classification. Poria et al. [20] extract the emotional features of text, speech and video and then concatenate them together and use the support vector machine (SVM) for emotion classification.

With the development of deep neural networks, multimodal sentiment analysis has yielded a series of results. Akhtar et al. [21] proposed a deep multitask learning framework to simultaneously learn emotional recognition and sentiment analysis in the multimodal context. Zhang et al. [22]

proposed integrating consistency and difference networks (ICDN) to map other modalities to the target modality through a cross-modal transformer. Then, unimodal labels are obtained by self-supervision for sentiment analysis tasks. Liu et al. [23] applied a low-rank multimodal fusion method to improve efficiency with a low-rank tensor. Experimental show that the method can reduce the number of parameters and improves the effectiveness of sentiment analysis. Yu et al. [24] designed a multilabel training program that can generate additional unimodal labels for each modality and perform training simultaneously with the main task.

Inspired by machine translation methods, some scholars have introduced the encoder-decoder structure into the field of multimodal sentiment analysis. The structure simulates human language by transforming the source modality into the target modality. Pham et al. [25] adopted a recurrent translation network to convert the original modality to the target modality and then convert the target modality back to the original modality to learn the joint embedding. Qi et al. [26] adopted bidirectional encoder representation from transformer (BERT) to encode multimodal data to resolve long-term dependencies within modalities. Mai et al. [27] used adversarial training to convert the distribution of source modalities to the distribution of target modalities. After the attention mechanism was proposed, the representation fusion method based on attention weighting developed rapidly. Li et al. [28] used interactive multi-head guided attention to capture the mapping relationship between multimodalities and map the results to higher dimensions through a soft-mapping layer. Yu et al. [29] proposed a one-way feature fusion method based on a multi-head attention mechanism to find correlations between modalities in different subspaces and locations.

Chen et al. [30] proposed a model based on SVM and CNN for emotion recognition. Miyazawa et al. [31] encode each modality using a unimodal pre-trained transformer model and fuses the unimodal encoded representations using a set of transformer layers. Huddar et al. [32] calculated the trimodal attention matrix based on the bimodal attention matrix to fuse the interaction information of different modalities. Xu et al. [33] proposed a coexistence network to capture the interaction information between modalities and then used the fused features for sentiment classification. Xi et al. [34] proposed a method based on multi-head attention mechanism, which adopted self-attention to capture modality features and used multi-head interactive attention to analyze the interactive information between different modalities.

Multimodal fusion is one of the challenge issues in multimodal sentiment analysis, which aims to integrate the information of multiple modalities and capture the interaction relationship between different modalities. Commonly used fusion methods can be divided into three types: feature-level fusion [5,22,28], decision-level fusion [14,35] and hybrid fusion [17,18]. For feature-level fusion, the commonly used method is to directly concatenate the features of each modality and construct a unified joint representation for prediction [36]. Although this method is simple and effective, it can only obtain shallow information and cannot deeply capture modality interactions, extract more abstract features, and generate input vectors that contain greater redundancy. For decision-level fusion, it usually makes independent predictions after obtaining the feature representation of each modality and obtaining the final decision result through weighting, majority voting, or deep neural network processing [37]. Decision-level fusion can make full use of the unique characteristics of each modality and has good generalization but ignores the correlation between the modalities.

The hybrid fusion strategy has the advantages of feature-level fusion and decision-level fusion. Nemati et al. [18] proposed a multimodal hybrid fusion method, which fuses audio and video using latent space, and then fuses projected features with text modalities using a dempster-shafer-based evidence fusion method. Cimtay et al. [38] proposed a multimodality-based emotion recognition model, which uses electroencephalogram (EEG) and galvanic skin response (GSR) modality feature fusion to estimate arousal levels, and then fuses EEG, GSR and facial modality features. However, existing hybrid fusion methods do not fuse the three modality features on the feature-level and decision-level and cannot learn the interactive information between multimodal data well. To improve the effect of multimodal sentiment analysis, we proposed an attention-based hybrid fusion strategy, which fuses three modalities on the feature-level and decision-level, the attention mechanism in the fusion strategy is useful for reducing redundant information and improving the prediction accuracy.

3 Methodology

To improve the accuracy of multimodal sentiment analysis, we propose a multimodal sentiment analysis model using BiGRU and attention-based hybrid fusion strategy (BAHFS). For the BAHFS model, the text (T), audio (A) and visual (V) are taken as the input. BAHFS consists of three parts, which are the unimodal representation learning part, feature-level fusion part and decision-level fusion part. The architecture of the model is shown in Fig. 1. For the unimodal representation learning part, we apply three independent BiGRU networks to learn the deep context features of the three unimodal data (T, V, A). For the feature-level fusion part, three fusion modules (including the bimodal attention fusion module, trimodal attention fusion module and trimodal concatenation fusion module) are proposed to conduct feature-level fusion. In the decision-level fusion part, the maximum rule is adopted to fuse the results obtained through classifying the two sets of the trimodal features and obtain the final sentiment analysis results. The calculation process of the BAHFS model will be introduced in detail as follows.

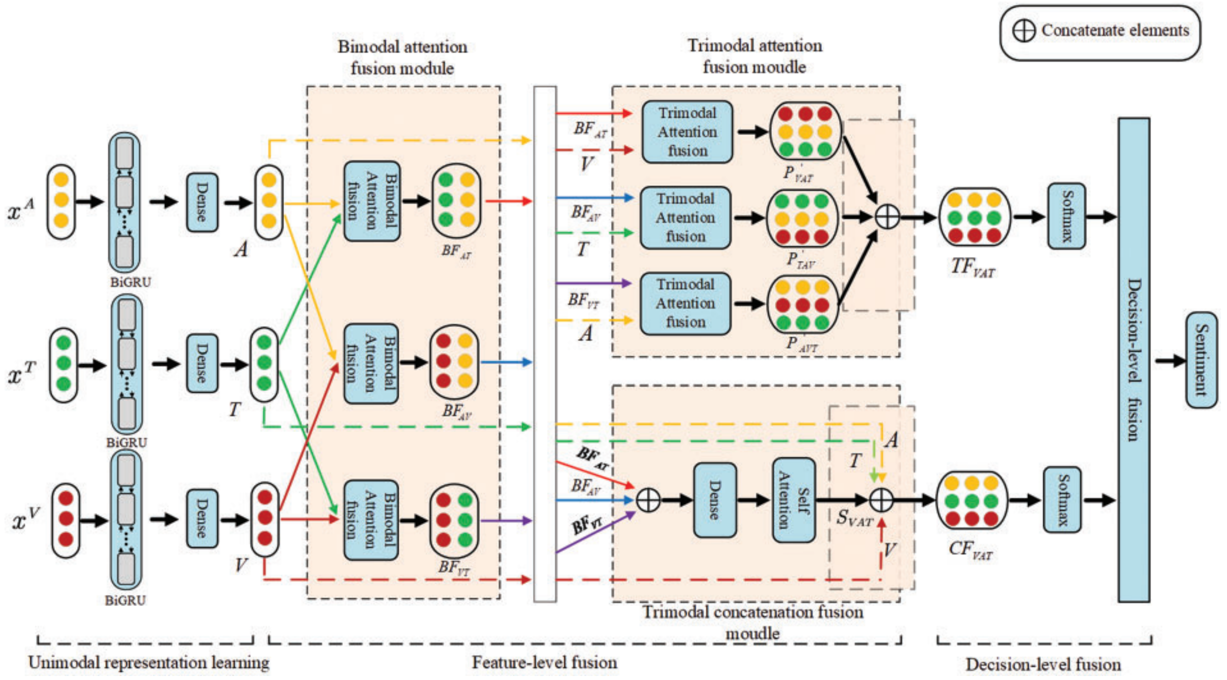

Figure 1: Architecture of the BAHFS model

3.1 Unimodal Representation Learning Part

Multimodal sentiment data are interdependent in terms of time series and semantics. Therefore, it needs to consider the context of the data in sentiment analysis. LSTM and gated recurrent units (GRU) are both recurrent neural network (RNN) and have equally excellent performance in some cases. Moreover, GRU requires fewer parameters than LSTM and can effectively overcome the overfitting problem. However, the GRU neural network only can access past information [39], to solve this problem, the BiGRU network is proposed, which can learn more information from both the past and the future using two hidden layers [40,41]. In this work, we adopt BiGRU to learn the deep contextual features of the unimodal data, the calculation process of BiGRU is described as Eqs. (1)–(3):

$$\overrightarrow{h}_t^m = GRU\left(x_t^m, \overrightarrow{h}_{t-1}^m\right) \tag{1}$$

$$\overleftarrow{h}_t^m = GRU\left(x_t^m, \overleftarrow{h}_{t+1}^m\right) \tag{2}$$

$$h_t^m = \overrightarrow{h}_t^m \oplus \overleftarrow{h}_t^m \tag{3}$$

where x_t^m indicates the t − th feature of input data, $\overrightarrow{h}_t^m$ means the forward hidden layer state of the t − th feature, $\overleftarrow{h}_t^m$ is the reverse hidden layer state of the t − th feature, h_t^m indicates the hidden layer state of BiGRU, and $\oplus$ denotes the concatenation operation.

Assuming that a video has u utterances, and the raw multimodal features of the three unimodal data are represented as follows: $x^{text} \in \mathbb{R}^{u\times d_T}$ denotes the raw textual features, $x^{audio} \in \mathbb{R}^{u\times d_A}$ means the raw audio features and $x^{visual} \in \mathbb{R}^{u\times d_V}$ represents the raw visual features. Firstly, the raw features of text, audio and visual are input into three BiGRU networks separately. Then, the deep context features of the multimodal data in the complete video sequence (denoted as X^{text}, X^{audio} and X^{visual}) are obtained. Next, the deep context features (X^{text}, X^{audio} and X^{visual}) are mapped to the same dimension with a fully connected layer and get the text features ($T \in \mathbb{R}^{u\times d}$), visual features ($V \in \mathbb{R}^{u\times d}$) and audio features ($A \in \mathbb{R}^{u\times d}$). Taking text modality as an example, the process of extracting the deep context features T is illustrated as Eqs. (4), (5):

$$X^{text} = BiGRU\left(x^{text}\right) \tag{4}$$

$$\mathrm{T} = \tanh\left(W_T X^{text} + b_T\right) \tag{5}$$

where BiGRU(·) denotes a bidirectional gated recurrent unit, W_T and b_T indicates the parameter matrix and bias of the fully connected layer, which is adopted to adjust the representations of the three modalities to the same dimension d. The audio context feature $A \in \mathbb{R}^{u\times d}$ and the visual context feature $V \in \mathbb{R}^{u\times d}$ can be obtained by the same method.

3.2 Features-Level Fusion Part

In this work, the features-level fusion includes bimodal features fusion and trimodal features fusion. Here we use the bimodal attention fusion module to realize the bimodal features fusion [11] and propose the trimodal attention fusion module and trimodal concatenation fusion module to realize the trimodal features fusion. In the following sections, we will introduce these modules in detail.

3.2.1 Bimodal Attention Fusion Module

To capture the interactive information between cross-modality and emphasize the important contributing features with the attention mechanism, based on the deep contextual features of the unimodal data, we carry out bimodal attention fusion on the unimodal features to obtain cross-modality (i.e., visual-text, text-audio and audio-visual) feature representations. Taking the text-visual for example, the bimodal attention fusion process is illustrated in Eq. (6):

$$M_1 = V \otimes T^{\mathrm{T}} \;\&\; M_2 = T \otimes V^{\mathrm{T}} M_1, M_2 \in \mathbb{R}^{u\times u} \tag{6}$$

where V and T are obtained from the BiGRU network in the unimodal representation learning part, V and T contain the contextual information of the text and visual modalities. T^{T} and V^{T} denote the transpose of the text and visual context feature matrices, respectively. $\otimes$ denotes a matrix product operation. $M_1, M_2 \in \mathbb{R}^{u\times u}$ are the bimodal attention matrices.

Then, we compute the probability distribution scores over each utterance of bimodal attention matrices M_1 and M_2 with a softmax function and get the attention distribution matrices ($N_1 \in \mathbb{R}^{u\times u}, N_2 \in \mathbb{R}^{u\times u}$), which can be illustrated in Eqs. (7) and (8). Finally, the matrix product is applied over the attention distribution matrices ($N_1 \in \mathbb{R}^{u\times u}, N_2 \in \mathbb{R}^{u\times u}$) to compute the modality-wise attentive representations $O_1 \in \mathbb{R}^{u\times d}$ and $O_2 \in \mathbb{R}^{u\times d}$, which can be described as Eq. (9).

$$N_1\,(p,q) = \frac{e^{M_1(p,q)}}{\sum_{k=1}^{u} e^{M_1(p,k)}} \quad \text{for } p,q = 1,\ldots,u \tag{7}$$

$$N_2\,(p,q) = \frac{e^{M_2(p,q)}}{\sum_{k=1}^{u} e^{M_2(p,k)}} \quad \text{for } p,q = 1,\ldots,u \tag{8}$$

$$O_1 = N_1 \otimes T,\; O_2 = N_2 \otimes V \tag{9}$$

where $N_1\,(p,q)$ denotes the correlation score of the $p-$th feature of the text modality and the $q-$th feature of the visual modality, which essentially represents the attention weights for the bimodal features. $\otimes$ denotes a matrix product operation. Next, an elementwise multiplication operation is carried out between the modality-wise attentive representations and the individual modality to obtain the interactive attention matrices, which are denoted as $A_1 \in \mathbb{R}^{u\times d}$ and $A_2 \in \mathbb{R}^{u\times d}$. Finally, the interactive attention matrices A_1 and A_2 are concatenated to obtain the fused text-visual bimodal feature $BF_{VT} \in \mathbb{R}^{u\times 2d}$. The above calculation process is shown as Eqs. (10), (11):

$$A_1 = O_1 \odot V,\; A_2 = O_2 \odot T \tag{10}$$

$$BF_{VT} = A_1 \oplus A_2 \tag{11}$$

where $\oplus$ denotes the concatenation operation, $\odot$ means the elementwise multiplication operation, which is useful for capturing the important information of multiple modalities. Through the same method, the fused text-audio BF_{AT} and visual-audio BF_{VA} bimodal features can be obtained.

3.2.2 Trimodal Attention Fusion Module

Since bimodal attention fusion can only capture the interaction information between two modalities and cannot capture the interaction information among the three modalities. To tackle this issue, we propose a trimodal attention fusion method to draw the deep interaction information among the three modalities. Here, we take the fusion between the bimodal feature BF_{VT} and unimodal audio A as an example to show the trimodal attention fusion process, which can be illustrated in Fig. 2.

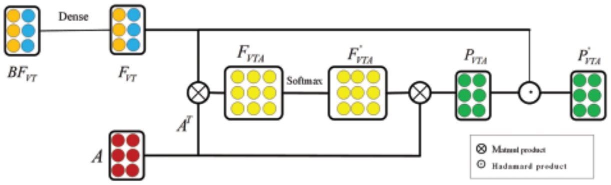

Figure 2: The framework of trimodal attention fusion

Firstly, we input the bimodal feature BF_{VT} into the fully connected layer to obtain the dimensionally reduced bimodal feature $F_{VT} \in \mathbb{R}^{u\times d}$ with the same dimension as the unimodal feature, this can be illustrated as Eq. (12). Then, the trimodal interaction matrix $F_{VTA} \in \mathbb{R}^{u\times u}$ is obtained through matrix product operation between the bimodal feature $F_{VT} \in \mathbb{R}^{u\times d}$ and the transposition of audio context feature matrix, which is illustrated as Eq. (13). After that, we compute the probability distribution scores over each utterance of trimodal interaction matrices F_{VTA} with a softmax function to obtain the trimodal probability distribution matrix $F'_{VTA} \in \mathbb{R}^{u\times u}$, which is illustrated as Eq. (14). Finally, the matrix product is applied over the trimodal probability distribution matrix F'_{VTA} to compute the trimodal attention representations $P_{VTA} \in \mathbb{R}^{u\times d}$, the calculation process is shown as Eq. (15):

$$F_{VT} = \tanh\left(W_v \times BF_{VT} + b_v\right) \tag{12}$$

$$F_{VTA} = F_{VT} \odot A^{\mathrm{T}} \tag{13}$$

$$F'_{VTA} = \mathrm{Softmax}\left(W_a \times F_{VTA} + b_a\right) \tag{14}$$

$$P_{VTA} = F'_{VTA} \otimes A \tag{15}$$

where $\tanh(\cdot)$ and $\mathrm{softmax}(\cdot)$ are the activation functions in the fully connected layer, W and b indicates the parameter matrices and bias of the fully connected layer, A^{T} is the transpose of the audio context feature matrices.

Next, an elementwise multiplication is computed between the trimodal attention matrix $P_{VTA} \in \mathbb{R}^{u\times d}$ and the bimodal feature matrix F_{VT} to obtain the trimodal interactive attention matrix $P'_{VTA} \in \mathbb{R}^{u\times d}$. With the same method, the trimodal interactive attention matrices P'_{ATV} and P'_{VAT} can be obtained. Finally, we concatenate the trimodal interactive attention matrices P'_{VTA}, P'_{VAT} and P'_{ATV} to produce the final trimodal features $TF_{VAT} \in \mathbb{R}^{u\times 3d}$. The calculation formulas for P'_{ATV} and TF_{VAT} are defined as Eqs. (16), (17):

$$P'_{VTA} = P_{VTA} \odot F_{VT} \tag{16}$$

$$TF_{VAT} = P'_{VTA} \oplus P'_{ATV} \oplus P'_{VAT} \tag{17}$$

where $\otimes$ denotes the matrix product operation, $\odot$ is the elementwise multiplication and $\oplus$ indicates a concatenation operation.

3.2.3 Trimodal Concatenation Fusion Module

To fully capture the connection between the trimodal features, we propose a trimodal concatenation fusion module for feature-level fusion. Firstly, we apply the concatenation operation on the three sets of bimodal features (visual-text, audio-text and audio-visual) and obtain a set of trimodal features, which is denoted as $B_{VAT} \in \mathbb{R}^{u\times 6d}$. The trimodal features contain all the bimodal feature information. Next, we reduce the dimensionality of the trimodal feature B_{VAT} through a fully connected layer and obtain C_{VAT}. Then, we apply the self-attention mechanism [11] to calculate the attention distribution of the trimodal feature C_{VAT} to get the trimodal feature S_{VAT}. Finally, we concatenate all unimodal features (visual features V, audio features A, text features T) and the trimodal features S_{VAT} to obtain another set of trimodal features $CF_{VAT} \in \mathbb{R}^{u \times 4d}$. The calculation formulas for B_{VAT}, C_{VAT}, S_{VAT} and CF_{VAT} are defined as Eqs. (18)–(21):

$$B_{VAT} = BF_{VT} \oplus BF_{AT} \oplus BF_{AV} \tag{18}$$

$$C_{VAT} = \tanh\left(W_c \times B_{VAT} + b_c\right) \tag{19}$$

$$S_{VAT} = \mathrm{SelfAttention}\left(C_{VAT}\right) \tag{20}$$

$$CF_{VAT} = S_{VAT} \oplus T \oplus V \oplus A \tag{21}$$

where W_c, b_c and $\tanh(\cdot)$ indicates the parameter matrix, bias and activation function of the fully connected layer, respectively.

3.3 Decision-Level Fusion Part

Decision-level fusion is high-level fusion, which can produce the final classification result by fusing the classification results of different modalities. The advantage of decision-level fusion is good anti-interference and fault tolerance. Currently, a variety of decision-level fusion methods have been proposed, including the sum rule, product rule, weighted average [14], maximum/minimum/median rule, majority voting [13], etc. In our work, the two sets of trimodal features TF_{VAT} and CF_{VAT} are obtained by using different feature-level fusion methods, and the sentiment probability scores of the two sets of trimodal fusion features can be obtained through the softmax function. Because the higher the probability score, the closer to the real sentiment. Therefore, we employ the maximum rule to select the highest score among two sets of trimodal features for decision-level fusion. Firstly, the softmax classifier is used to classify the two sets of trimodal features TF_{VAT} and CF_{VAT} respectively, the output of each classifier is regarded as a classification score, which can be illustrated as Eqs. (22) and (23). Then, the decision-level fusion is conducted on the classification results of the two sets of trimodal features, the calculation formula is defined as Eq. (24).

$$l_1 = \mathrm{softmax}\left(W_1 \times CF_{VAT} + b_1\right) \tag{22}$$

$$l_2 = \mathrm{softmax}\left(W_2 \times TF_{VAT} + b_2\right) \tag{23}$$

$$\hat{L} = \mathrm{maximum}\left(l_1, l_2\right) \tag{24}$$

where l_1 and l_2 are the classification probability scores calculated by the softmax function on the two sets of trimodal features. $W_{1,2}$ and $b_{1,2}$ are the weights and biases of the softmax classification layer, and $\hat{L}$ is the final sentiment category. The calculation process of our proposed multimodal sentiment analysis model is illustrated as Algorithm 1.

Algorithm 1: Multimodal sentiment analysis model based on BiGRU and hybrid fusion strategy

```
Input:
$x^m \in \mathbb{R}^{u \times d_m}, m \in \{text, audio, visual\}$: The raw multimodal features.
Output:
$L$: The final sentiment category.
Phase I. Deep Context Feature Extraction
    $X^m \leftarrow BiGRU(x^m), m \in \{text, audio, visual\}$
```

(Continued)

Algorithm 1: Continued

```
  $T \leftarrow \tanh(W_T \times X^{text} + b_T)$
  $return\,(T)$
Phase II. Bimodal attention fusion
  Calculate a pair of bimodal attention matrices $M_1$, $M_2$ according to Eq. (6).
  Compute the probability distribution scores $(N_1, N_2)$ over each utterance of $M_1$,$M_2$ according to Eqs. (7)–(9).
  Compute the interactive attention matrices $A_1$, $A_2$ according to Eq. (10) and concatenate them.
  $BF_{VT} \leftarrow$ Bimodalattention fusion $(V, T)$
  $BF_{VA} \leftarrow$ Bimodalattention fusion $(V, A)$
  $BF_{AT} \leftarrow$ Bimodalattention fusion $(A, T)$
  $return\,(BF_{VT}, BF_{VA}, BF_{AT})$
Phase III. Trimodal attention fusion
  Generate the trimodal features $P'_{VTA}$, $P'_{ATV}$, $P'_{VAT}$ according to Eqs. (12)–(15).
  $TF_{VAT} \leftarrow Concatenate\left(P'_{VTA}, P'_{ATV}, P'_{VAT}\right)$
  $return\,(TF_{VAT})$
Phase IV. Trimodal concatenation fusion
  Generate the bimodal features $BF_{VT}$, $BF_{VA}$, $BF_{AT}$ according to Phase II.
  $B_{VAT} \leftarrow Concatenate\,(BF_{VT}, BF_{VA}, BF_{AT})$
  Produce $S_{VAT}$ from $B_{VAT}$ according to Eqs. (19), (20)
  $CF_{VAT} \leftarrow Concatenate\,(S_{VAT}, V, A, T)$
  $return\,(CF_{VAT})$
Phase V. Decision-level fusion
  Get $TF_{VAT}$ and $CF_{VAT}$ from Phase III and IV.
  $l_1 \leftarrow softmax\,(TF_{VAT})$
  $l_2 \leftarrow softmax\,(CF_{VAT})$
  $\hat{L} \leftarrow$ maxmium $(l_1, l_2)$
  $return\,\hat{L}$
End
```

3.4 Training

The goal of model BAHFS is to perform multimodal sentiment analysis whose predictions are given by $\hat{L}$, here we use cross-entropy to measure the loss between the predictions and true labels. Given the true labels as L, the sentiment analysis loss is computed as Eq. (25).

$$\varphi = -\frac{1}{N}\sum_{i=1}^{N}\hat{L}_i \cdot \log L_i + (1 - L_i) \cdot \log\left(1 - \widehat{L_i}\right) \tag{25}$$

where φ is the loss for the sentiment analysis task. L and $\hat{L}$ are the true label and the predicted label, respectively. In the BAHFS model, the Adam optimizer is used to optimize the model parameters.

4 Experiments

To verify the performance of our proposed model BAHFS, we take the public datasets CMU-MOSI [42] and CMU-MOSEI [43] as the testing data, which includes the audio, text, and video modalities. The experimental environment is configured as follows. OS: Windows 10, CPU: Intel

Core i9-10900 K, GPU: GeForce RTX 3090. The BAHFS model is implemented with the Keras deep learning architecture supported by the backend of TensorFlow 2.6. Representative multimodal sentiment analysis models are selected as the comparison models, and the accuracy score and F1 score are used as the evaluation metrics. We will introduce our experiment in detail as follows.

4.1 Datasets

CMU-MOSI contains 93 videos with a total of 2199 utterances. We take 52 videos (1151 utterances) as the training set, take 10 videos (297 utterances) as the validation set, and use 31 videos (752 utterances) as the test set. CMU-MOSEI is a large dataset that contains 3229 videos with a total of 22676 utterances. We take 2250 videos (16216 utterances) as the training set, take 300 videos (1835 utterances) as the validation set, and take 679 videos (4625 utterances) as the test set. The specific information on the CMU-MOSI dataset and the CMU-MOSEI dataset is shown in Table 1.

Table 1: Datasets statistics in CMU-MOSI and CMU-MOSEI

Datasets	Train	Valid	Test	Total
CMU-MOSI	1151	296	752	2199
CMU-MOSEI	16216	1835	4625	22676

Each utterance in the CMU-MOSI dataset is labeled as positive or negative, while in the CMU-MOSEI dataset, the labels are set in the range of -3 to $+3$. To compare with the existing approaches, we project the CMU-MOSEI instance into a two-class classification setting, where "values ≥ 0" indicates the positive sentiments and "values < 0" indicates the negative sentiments.

4.2 Feature Extraction

To obtain reliable feature input and load data efficiently, for the CMU-MOSI dataset, we use utterance-level features provided in [44]. These features are extracted with a CNN [45], 3D CNN [46] and openSMILE [47] for text, visual and audio modalities, respectively. The embedding dimensions for text, audio and visual features are 100, 100, and 73, respectively. For the CMU-MOSEI dataset, we use the CMU-MOSEI dataset provided by Poria et al. [44]. The text features of this dataset are extracted by GloVe embedding, visual features are extracted by Facets, and audio features are extracted by CovaRep [48]. The dimensions of the feature vector are set to 300 (text), 35 (visual) and 74 (audio).

4.3 Parameters Setting and Evaluation Standard

For the CMU-MOSI dataset, the number of neurons of BiGRU is set as 300 and set dropout = 0.7. For the CMU-MOSEI dataset, the number of neurons of BiGRU is set as 200 and set dropout = 0.5. Each BiGRU is followed by a fully connected layer consisting of 100 neurons. We project the features of the three modalities to the same dimension through a fully connected layer. For the fully connected layer, the tanh activation function is adopted, and we set dropout = 0.5 for the CMU-MOSI dataset and set dropout = 0.3 for the CMU-MOSEI dataset. For the classification layer, the softmax activation function is adopted. When training the BAHFS model, the batch size is set to 32 and the number of iterations is set to 50.

4.4 Baselines

To verify the performance of our model, 11 representative multimodal sentiment analysis models are selected as the baseline models. The details of the 11 baseline models are introduced as follows:

- DFF-ATMF [36]: proposed a fusion strategy combining multi-feature fusion and multimodal fusion to improve the accuracy of sentiment analysis.
- ICDN [22]: proposed the integrating consistency and difference networks (ICDN) to map other modalities to the target modality through a cross-modal transformer. Then, unimodal sentiment labels are obtained by self-supervision for sentiment analysis tasks.
- MCTN [25]: Use a Seq2Seq model to learn multimodal representations by translating between modalities.
- MFN [10]: Use gated memory cells to store the internal information of modalities and the interaction information between modalities and add dynamic fusion graphs to represent emotional information.
- RAVEN [13]: Model the nonverbal sub-word sequences and dynamically adjust word representations based on nonverbal cues.
- Graph-MFN [43]: proposed a dynamic fusion graph to analyze the interactions between different modalities and use LSTM to capture the overall information sequence.
- AWMA [49]: proposed a multi-attention neural network called an asymmetric window, using sliding asymmetric windows to represent different weights of historical and future contexts at specific timestamps of input data in different modes.
- MMMU-BA [11]: proposed a BiGRU-based multimodal attention framework, which extracts contextual information for utterance-level sentiment analysis.
- Multiogue-Net [50]: proposed an RNN architecture for multimodal sentiment analysis in conversation. The model can effectively capture the contextual information of the dialogue while considering the information saved by a specific modality.
- MulT [51]: proposed a multimodal transformer (MulT) to analyze human multimodal language. The model utilizes a cross-modal attention mechanism to map latent information from one modality to another modality.
- TCM-LSTM [52]: proposed a temporal convolutional multimodal LSTM model, which is dominated by linguistic features and assisted by audio features and visual features and uses LSTM with a gating mechanism to learn cross-modal information from different perspectives.

We test the proposed BAHFS model against 11 baseline models on the CMU-MOSI and CMU-MOSEI datasets. For the two datasets, the results of the 11 baseline models are obtained from the original paper, and the results of the BAHFS model are obtained by testing it on our experimental platform. For the small dataset CMU-MOSI, experimental results are presented in Table 2. From Table 2, we can find that the BAHFS model ranks second in terms of accuracy, and the accuracy of model BAHFS is 0.13% lower than that of the MMMU-BA model, the accuracy of the BAHFS model is better than that of the other 10 baseline models. Moreover, BAHFS gets the best result on the F1 metric among the 12 comparison models.

For the large dataset CMU-MOSEI, the BAHFS model gets the best results in terms of accuracy and F1 score among the 12 models. Specifically, compared with the Multiogue-Net model [50] which has the second highest accuracy, the accuracy of the BAHFS model is 2.53% larger than that of the Multiogue-Net model, and the F1 score of the BAHFS model is 3.77% larger than that of Multiogue-Net model. Compared with the MulT [51] model which has the second highest F1 score, the BAHFS model improves the accuracy by 3.03% and improves the F1 score by 2.18%. Compared with the MMMU-BA model [11], which is most like BAHFS, the accuracy of BAHFS is 4.83% larger than that of MMMU-BA. The above experimental results prove that the overall performance of the BAHFS model is better than that of the other 11 baseline models on the large dataset CMU-MOSEI.

Table 2: Performance on CMU-MOSI and CMU-MOSEI datasets

Modality	CMU-MOSEI		CMU-MOSI	
	Acc	F1 score	Acc	F1 score
DFF-ATMF	77.10	78.30	80.98	81.26
ICDN	82.70	82.50	81.50	81.60
MCTN	–	–	79.30	79.10
MFN	–	–	77.40	77.30
RAVEN	79.10	79.50	78.00	76.60
Graph-MFN	76.90	77.00	–	–
AWMA	–	–	80.00	79.90
MMMU-BA	79.80	–	**82.31**	–
Multiogue-Net	82.10	80.01	81.19	80.01
MulT	81.60	81.60	81.10	81.00
TCM-LSTM	81.40	81.60	81.70	81.80
BAHFS	**84.63**	**83.78**	82.18	**81.96**

Compared with the MMMU-BA model [11], the BAHFS model adds many modules, which also leads to a slight increase in the complexity of the model. BAHFS runs an epoch on the CMU-MOSI dataset for 4 s, which is no different from the running time of the MMMU-BA model. The running time of an epoch on the CMU-MOSEI dataset is 55 s, which is 7 s more than the 48 s of the MMMU-BA model, the parameter size of our model is 30 M.

4.5 Ablation Experiments

The ablation experiments include modality ablation experiments and model ablation experiments. The modality ablation experiments aim to show the performance of the BAHFS model with unimodal, bimodal and trimodal data. The model ablation experiments aim to verify the effect of different fusion strategies in the BAHFS model. Moreover, through the ablation experiments, the effectiveness of the BAHFS model is verified.

4.5.1 Modality Ablation Experiments

Existing research work [53,54] proves that multimodal data contain richer information than unimodal data and are useful for revealing people's sentiments. To verify the influence of different multimodal data on sentiment analysis, we take the unimodal data features (visual feature V, audio feature A and text feature T), the bimodal feature combination (V+A, V+T and A+T) and trimodal feature combination (V+A+T) as the input of the BAHFS model, and then carry out the sentiment analysis experiment as follows.

(i) We first input the unimodal features into the BiGRU network to learn the context features. Then input the output of BiGRU into the fully connected layer for dimensionality reduction. Finally, the softmax function is used to output the sentiment analysis result. (ii) We combine the unimodal feature (V, A, T) to obtain different bimodal features (V+A, V+T and A+T). Then we use the bimodal attention fusion mechanism [11] to perform the bimodal feature fusion, and then input the bimodal fusion features into the fully connected layer with tanh activation function. Finally, the softmax classifier is adopted for sentiment analysis. (iii) We input the trimodal features into BAHFS model and get the sentiment analysis result. The experimental results of the modality ablation experiment are presented in Table 3.

Table 3: Experimental results of modality ablation on CMU-MOSI and CMU-MOSEI datasets

Modality	CMU-MOSEI		CMU-MOSI	
	Acc	F1 score	Acc	F1 score
V	81.59	74.66	60.11	54.53
A	81.96	76.29	62.23	62.19
T	83.18	82.20	79.65	77.17
V+A	82.94	82.27	64.76	58.82
V+T	83.35	83.34	80.32	78.75
A+T	83.44	82.77	80.45	79.02
V+A+T	**84.63**	**83.78**	**82.18**	**81.96**

From Table 3 we can find that, for the unimodal sentiment analysis, on the CMU-MOSEI and CMU-MOSI datasets, the accuracy and F1 score of sentiment analysis based on text modality are the highest, which are (79.65%, 77.17%) and (83.18%, 82.20%), respectively. Moreover, the accuracy and F1 score of the text modality are 17.42% and 14.98% larger than those of the audio modality on the CMU-MOSI dataset, and 1.22% and 5.91% larger than those of the audio modality on the CMU-MOSEI dataset. Compared with the visual modality, the accuracy and F1 score of the text modality are 19.54% and 22.64% larger than those of the visual modality on the CMU-MOSI dataset and 1.59% and 7.57% larger than those of the visual modality on the CMU-MOSEI dataset. The experimental results demonstrates that the text modality dominates the unimodal sentiment analysis task. This is because the textual modality contains more sentiment features than the other two modalities. Experimental results also can prove that the text modality contains less redundant information than that of the audio and visual modalities, which is more conducive to identifying real sentiment.

From Table 3 we can see that the results of the bimodal (A+V) are much lower than those of the bimodal with text modality, and the bimodal (A+T) has the best results. The accuracy and F1 score of the bimodal (A+T) are 80.45% and 79.02% on the CMU-MOSI dataset, and 83.44% and 82.77% on the CMU-MOSEI dataset. The accuracy of the bimodal feature combination (A+T) is 15.69% (on the CMU-MOSI dataset) and 0.5% (on the CMU-MOSEI dataset) larger than that of the bimodal feature combination (A+V). The above results proves that the combination of the bimodal features including the text modality has more valuable information for sentiment analysis.

From Table 3 we can find that the sentiment analysis model BAHFS can achieve the best performance with the trimodal features on the CMU-MOSI and CMU-MOSEI datasets. Specifically, the accuracy and F1 score of the BAHFS model on the CMU-MOSI dataset are 82.18% and 81.96%, and the accuracy and F1 score of the BAHFS model on the CMU-MOESI dataset are 84.63% and 83.78%. These experimental results proves that the sentiment analysis model with trimodal features can effectively reflect the real sentiment of users.

4.5.2 Model Ablation Experiment

To verify the effectiveness of the multimodal feature fusion method and hybrid fusion strategy in our model, an ablation experiment has been carried out. In this experiment, some new models are generated by changing or removing some operations or modules from BAHFS. The newly generated models are described as follows:

- BAHFS(BAF): "BAF" indicates the bimodal attention fusion module used in this work. BAHFS(BAF) is generated by removing the trimodal concatenation fusion module and the trimodal attention fusion module from BAHFS and keeping the bimodal attention fusion module. In this model, all the bimodal features obtained by the bimodal attention fusion module are concatenated.
- BAHFS(TCF): "TCF" presents the trimodal concatenation fusion module proposed in this work. BAHFS(TCF) is generated by removing the trimodal attention fusion module from BAHFS. In model BAHFS(TCF), the bimodal features obtained from the bimodal attention fusion module are concatenated. After the important features are captured by the self-attention mechanism, they are concatenated together with unimodal features into trimodal features for sentiment analysis.
- BAHFS(TAF): "TAF" means the trimodal attention fusion module proposed in this paper. BAHFS(TAF) is produced by removing the trimodal concatenation fusion module from BAHFS. In model BAHFS(TAF), the bimodal features and unimodal features fused by the bimodal attention fusion module are input into the trimodal attention fusion module, and all the obtained trimodal features are concatenated for sentiment analysis.
- BAHFS(DLF): "DLF" denotes the decision-level fusion adopted in this work. Model BAHFS(DLF) is obtained by removing feature-level fusion (including bimodal attention fusion module, trimodal attention fusion module and trimodal concatenation fusion module) from BAHFS. In BAHFS(DLF), the three unimodal depth context features are analyzed separately, and then decision-level fusion is performed.

To verify the effectiveness of our proposed trimodal attention fusion module, we compare the performance of model BAHFS(BAF) and BAHFS(TAF), where the bimodal attention fusion module is used in BAHFS(BAF) and trimodal attention fusion module is used in BAHFS(TAF). Experimental results are presented in Table 4. From Table 4 we can find that the accuracy and F1 score of BAHFS(TAF) in the CMU-MOSEI dataset are 0.68% and 0.76% higher than BAHFS(BAF), respectively. In the CMU-MOSI dataset, the accuracy and F1 score of BAHFS(TAF) are 0.33% and 0.61% higher than BAHFS(BAF), respectively. That is, BAHFS(TAF) achieves the best results in accuracy and F1 score, which can prove that our proposed trimodal attention fusion module is better than the bimodal attention fusion module.

To prove the effectiveness of our proposed hybrid fusion strategy, we compare the performance of model BAHFS(TCF), BAHFS(TAF), BAHFS(DLF) and BAHFS, where trimodal concatenation fusion module is adopted in BAHFS(TCF), trimodal attention fusion module is applied in BAHFS(TAF), decision-level fusion is used in BAHFS(DLF) and BAHFS. Experimental results are presented in Table 4.

From Table 4 we can find that when the trimodal attention fusion module and the trimodal concatenation fusion module are removed, that is BAHFS (BAF) model. Compared with the proposed BAHFS model, the accuracy and F1 score in the CMU-MOSEI dataset decreased by 1.61% and 1.48% and the accuracy and F1 score in the CMU-MOSI dataset decreased by 2.32% and 3.96%, respectively. When the trimodal attention fusion module is removed separately, that is BAHFS (TCF) model.

Compared with the proposed BAHFS model, the accuracy and F1 score in the CMU-MOSEI dataset decreased by 1.21% and 0.32%, and the accuracy and F1 score in the CMU-MOSI dataset decreased by 1.46% and 2.83%, respectively. When the trimodal concatenation fusion module is removed separately, that is BAHFS (TAF) model. Compared with the proposed BAHFS model, the accuracy and F1 score in the CMU-MOSEI dataset decreased by 0.93% and 0.72%, and the accuracy and F1 score in the CMU-MOSI dataset decreased by 1.99% and 3.35%, respectively. When the feature-level fusion part is removed, only decision-level fusion is applied, that is BAHFS (DLF) model. Compared with the proposed BAHFS model, the accuracy and F1 score in the CMU-MOSEI dataset decreased by 1.75% and 5.63%, and the accuracy and F1 score in the CMU-MOSI dataset decreased by 17.55% and 25.5%, respectively. Therefore, the BAHFS model has the best performance among the three models. These experimental results prove that our proposed hybrid fusion strategy is effective.

Table 4: Experimental results of modality ablation on CMU-MOSI and CMU-MOSEI datasets

Tasks	CMU-MOSI		CMU-MOSEI	
	Acc	F1 score	Acc	F1 score
BAHFS(BAF)	83.02	82.30	79.86	78.00
BAHFS(TCF)	83.42	83.46	80.72	79.13
BAHFS(TAF)	83.70	83.06	80.19	78.61
BAHFS(DLF)	82.88	78.15	64.63	56.46
BAHFS	84.63	83.78	82.18	81.96

5 Conclusion

To improve the performance of multimodal sentiment analysis, we propose a new multimodal sentiment analysis model using BiGRU and attention-based hybrid fusion Strategy (named BAHFS). In BAHFS, BiGRU is first used to capture the unimodal features from the unimodal data. Next, the unimodal features are fused into bimodal features through the bimodal attention fusion module. Then, the bimodal and unimodal features are sent into the trimodal concatenation fusion module and the trimodal attention fusion module to get two sets of trimodal features. Finally, the two sets of trimodal features are fused through the decision-level fusion strategy to obtain the final sentiment category. Experimental results shows that our proposed trimodal concatenation fusion module and hybrid fusion strategy are useful for enhancing the performance of multimodal sentiment analysis, and our proposed model has better performance than the other eleven baseline models. For the future work, we will investigate the multimodal sentiment analysis problem with uncertain missing modalities.

Funding Statement: This research was funded by the National Natural Science Foundation of China (Grant No. 61872126, No. 62273290). It was also supported by the Key project of Natural Science Foundation of Shandong Province (Grant No. ZR2020KF019).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] 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.

[2] C. P. Selvi, P. Muneeshwari, K. Selvasheela and D. Prasanna, "Twitter media sentiment analysis to convert non-informative to informative using QER," *Intelligent Automation & Soft Computing*, vol. 35, no. 3, pp. 3545–3555, 2023.

[3] L. Y. Qi, Y. H. Yang, X. K. Zhou, W. Rafique and J. H. Ma, "Fast anomaly identification based on multi-aspect data streams for intelligent intrusion detection toward secure industry 4.0," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 9, pp. 6503–6511, 2021.

[4] C. Peng, C. X. Zhang, X. J. Xue, J. M. Gao, H. J. Liang *et al.,* "Cross-modal complementary network with hierarchical fusion for multimodal sentiment classification," *Tsinghua Science and Technology*, vol. 27, no. 4, pp. 664–679, 2021.

[5] J. H. Xi, K. Q. Zheng, Y. F. Zhong, L. J. Li, Z. P. Cai *et al.,* "Robust symmetry prediction with multimodal feature fusion for partial shapes," *Intelligent Automation & Soft Computing*, vol. 35, no. 3, pp. 3099–3111, 2023.

[6] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[7] C. Y. Liao, R. C. Chen and S. K. Tai, "Emotion stress detection using eeg signal and deep learning technologies," in *2018 IEEE Int. Conf. on Applied System Invention (ICASI)*, Chiba, Japan, pp. 90–93, 2018.

[8] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L. P. Morency, "Deep multimodal fusion for persuasiveness prediction," in *Proc. of the 18th ACM Int. Conf. on Multimodal Interaction*, Tokyo, Japan, pp. 284–288, 2016.

[9] D. D. Deng, Y. Q. Zhou, J. M. Pi and B. E. Shi, "Multimodal utterance-level affect analysis using visual, audio and text features," arXiv preprint arXiv: 1805.00625, 2018.

[10] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria *et al.,* "Memory fusion network for multi-view sequential learning," in *Proc. of the AAAI Conf. on Artificial Intelligence*, New Orleans, LA, USA, vol. 32, no. 1, 2018.

[11] D. Ghosal, M. S. Akhtar, D. Chauhan, S. Poria, A. Ekbal *et al.,* "Contextual inter-modal attention for multimodal sentiment analysis," in *Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing*, Brussels, Belgium, pp. 3454–3466, 2018.

[12] R. R. Sarvestani and R. Boostani, "Ff-skpcca: Kernel probabilistic canonical correlation analysis," *Applied Intelligence*, vol. 46, no. 2, pp. 438–454, 2017.

[13] Y. S. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh *et al.,* "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proc. of the AAAI Conf. on Artificial Intelligence*, Honolulu, HI, USA, vol. 33, no. 1, pp. 7216–7223, 2019.

[14] M. H. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh *et al.,* "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proc. of the 19th ACM Int. Conf. on Multimodal Interaction*, Glasgow, UK, pp. 163–171, 2017.

[15] H. H. Wang, A. Meghawat, L. P. Morency and E. P. Xing, "Select-additive learning: Improving generalization in multimodal sentiment analysis," in *2017 IEEE Int. Conf. on Multimedia and Expo (ICME)*, Hong Kong, China, pp. 949–954, 2017.

[16] Y. H. Yu, H. F. Lin, J. N. Meng and Z. H. Zhao, "Visual and textual sentiment analysis of a microblog using deep convolutional neural networks," *Algorithms*, vol. 9, no. 2, pp. 41, 2016.

[17] H. Gunes and M. Piccardi, "Bimodal emotion recognition from expressive face and body gestures," *Journal of Network and Computer Applications*, vol. 30, no. 4, pp. 1334–1345, 2007.

[18] S. Nemati, R. Rohani, M. E. Basiri, M. Abdar, N. Y. Yen *et al.,* "A hybrid latent space data fusion method for multimodal emotion recognition," *IEEE Access*, vol. 7, pp. 172948–172964, 2019.

[19] D. L. Cao, R. R. Ji, D. Z. Lin and S. Z. Li, "A cross-media public sentiment analysis system for microblog," *Multimedia Systems*, vol. 22, pp. 4, pp. 479–486, 2016.

[20] S. Poria, E. Cambria, A. Hussain and G. B. Huang, "Towards an intelligent framework for multimodal affective data analysis," *Neural Networks*, vol. 63, pp. 104–116, 2015.

[21] M. S. Akhtar, D. S. Chauhan, D. Ghosal, S. Poria, A. Ekbal *et al.,* "Multi-task learning for multimodal emotion recognition and sentiment analysis," arXiv preprint arXiv: 1905.05812, 2019.

[22] Q. A. Zhang, L. Shi, P. Y. Liu, Z. F. Zhu and L. C. Xu, "Icdn: Integrating consistency and difference networks by transformer for multimodal sentiment analysis," *Applied Intelligence*, pp. 1–14, 2022.

[23] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh *et al.,* "Efficient low-rank multimodal fusion with modality-specific factors," arXiv preprint arXiv: 1806.00064, 2018.

[24] W. M. Yu, H. Xu, Z. Q. Yuan and J. L. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," arXiv preprint arXiv: 2102.04830, 2021.

[25] H. Pham, P. P. Liang, T. Manzini, L. P. Morency and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proc. of the AAAI Conf. on Artificial Intelligence*, Honolulu, HI, USA, vol. 33, pp. 6892–6899, 2019.

[26] Q. F. Qi, L. Y. Lin, R. Zhang and C. R. Xue, "Medt: Using multimodal encoding-decoding network as in transformer for multimodal sentiment analysis," *IEEE Access*, vol. 10, pp. 28750–28759, 2022.

[27] S. J. Mai, H. F. Hu and S. L. Xing, "Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion," in *Proc. of the AAAI Conf. on Artificial Intelligence*, New York, USA, vol. 34, no. 1, pp. 164–172, 2020.

[28] Z. H. Li, Q. B. Guo, C. Y. Feng, L. J. Deng, Q. W. Zhang *et al.,* "Multimodal sentiment analysis based on interactive transformer and soft mapping," *Wireless Communications and Mobile Computing*, vol. 2022, 2022.

[29] J. F. Yu and J. Jiang, "Adapting bert for target-oriented multimodal sentiment classification," in *Proc. of the Twenty-Eighth Int. Joint Conf. on Artificial Intelligence*, Macao, China, pp. 5408–5414, 2019.

[30] Y. L. Chen and Z. Zhang, "Research on text sentiment analysis based on cnns and svm," in *2018 13th IEEE Conf. on Industrial Electronics and Applications (ICIEA)*, Wuhan, China, pp. 2731–2734, 2018.

[31] K. Miyazawa, Y. Kyuragi and T. Nagai, "Simple and effective multimodal learning based on pre-trained transformer models," *IEEE Access*, vol. 10, pp. 29821–29833, 2022.

[32] 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.

[33] N. Xu, W. J. Mao and G. D. Chen, "A co-memory network for multimodal sentiment analysis," in *The 41st int. ACM SIGIR Conf. on Research & Development in Information Retrieval*, Ann Arbor, MI, USA, pp. 929–932, 2018.

[34] C. Xi, G. M. Lu and J. J. Yan, "Multimodal sentiment analysis based on multi-head attention mechanism," in *Proc. of the 4th Int. Conf. on Machine Learning and Soft Computing*, Haiphong City, Viet Nam, pp. 34–39, 2020.

[35] S. Koelstra, C. Muhl, M. Soleymani, J. S. Lee, A. Yazdani *et al.,* "Deap: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2011.

[36] F. Y. Chen, Z. Q. Luo, Y. Y. Xu and D. F. Ke, "Complementary fusion of multi-features and multimodalities in sentiment analysis," arXiv preprint arXiv: 1904.08138, 2019.

[37] S. A. Abdu, A. H. Yousef and A. Salem, "Multimodal video sentiment analysis using deep learning approaches, a survey," *Information Fusion*, vol. 76, pp. 204–226, 2021.

[38] Y. Cimtay, E. Ekmekcioglu and S. Caglar-Ozhan, "Cross-subject multimodal emotion recognition based on hybrid fusion," *IEEE Access*, vol. 8, pp. 168865–168878, 2020.

[39] Y. P. Deng, H. Jia, P. C. Li, X. Q. Tong, X. D. Qiu *et al.,* "A deep learning methodology based on bidirectional gated recurrent unit for wind power prediction," in *2019 14th IEEE Conf. on Industrial Electronics and Applications (ICIEA)*, Xi'an, China, pp. 591–595, 2019.

[40] X. Luo, W. W. Zhou, W. P. Wang, Y. Q. Zhu and J. Deng, “Attention-based relation extraction with bidirectional gated recurrent unit and highway network in the analysis of geological data,” *IEEE Access*, vol. 6, pp. 5705–5715, 2017.

[41] D. J. Zhang, L. Tian, M. B. Hong, F. Han, Y. F. Ren *et al.,* “Combining convolution neural network and bidirectional gated recurrent unit for sentence semantic classification,” *IEEE Access*, vol. 6, pp. 73750–73759, 2018.

[42] A. Zadeh, R. Zellers, E. Pincus, and L. P. Morency, “Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages,” *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[43] A. Zadeh and P. Pu, “Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph,” in *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, pp. 2236–2246, 2018.

[44] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh *et al.,* “Context-dependent sentiment analysis in user-generated videos,” in *Proc. of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, pp. 873–883, 2017.

[45] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar *et al.,* “Large-scale video classification with convolutional neural networks,” in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Columbus, OH, USA, pp. 1725–1732, 2014.

[46] S. W. 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, 2012.

[47] F. Eyben, M. Wöllmer and B. Schuller, “Opensmile: The munich versatile and fast open-source audio feature extractor,” in *Proc. of the 18th ACM Int. Conf. on Multimedia*, Firenze, Italy, pp. 1459–1462, 2010.

[48] G. Degottex, J. Kane, T. Drugman, T. Raitio and S. Scherer, “Covarep—A collaborative voice analysis repository for speech technologies,” in *2014 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, pp. 960–964, 2014.

[49] H. L. Lai and X. M. Yan, “Multimodal sentiment analysis with asymmetric window multi-attentions,” *Multimedia Tools and Applications*, vol. 81, no. 14, pp. 19415–19428, 2022.

[50] A. Shenoy and A. Sardana, “Multilogue-net: A context aware rnn for multimodal emotion detection and sentiment analysis in conversation,” arXiv preprint arXiv: 2002.08267, 2020.

[51] Y. H. H. Tsai, S. J. Bai, P. P. Liang, J. Z. Kolter, L. P. Morency *et al.,* “Multimodal transformer for unaligned multimodal language sequences,” in *Proc. of the Conf. Association for Computational Linguistics. Meeting*, Florence, Italy, vol. 2019, pp. 6558, 2019.

[52] S. J. Mai, S. L. Xing and H. F. Hu, “Analyzing multimodal sentiment via acoustic-and visual-lstm with channel-aware temporal convolution network,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1424–1437, 2021.

[53] L. Y. Qi, W. M. Lin, X. Y. Zhang, W. C. Dou, X. L. Xu *et al.,* “A correlation graph based approach for personalized and compatible web apis recommendation in mobile app development,” *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[54] V. P. Rosas, R. Mihalcea and L. P. Morency, “Multimodal sentiment analysis of spanish online videos,” *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 38–45, 2013.