Pattern Recognition 156 (2024) 110847

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Sentiment analysis based on text information enhancement and multimodal feature fusion

Zijun Liu, Li Cai *, Wenjie Yang, Junhui Liu

School of Software, Yunnan University, Kunming, 650500, China

ARTICLE INFO

Keywords:
Sentiment analysis
Text information enhancement
Multimodal data fusion
Cross-modal attention mechanism
Sentiment lexicons

ABSTRACT

Rapid advancements in multimedia technology have created explosive growth in sentiment data generated across various social media platforms. While previous research on sentiment analysis has shifted from analyzing single data types to incorporating multimodal data, current studies face certain limitations. These include overlooking the impact of redundant information within feature sequences of each modality, failing to account for the complementarity between modality data, and neglecting the varying significance of different modalities in conveying sentiments. This paper introduces a sentiment analysis framework designed for text information enhancement and multimodal feature fusion. The text modality is central to this framework, around which an attention mechanism augments emotional correlations between modalities. An expanded sentiment lexicon refines the representation of multimodal features, thus capturing emotional information more accurately. Experimental evaluations conducted on two standard datasets, CMU-MOSI and CMU-MOSEI, show that the accuracy of the proposed method in multimodal emotion recognition tasks reaches 85.7% and 85.8% respectively, at 1.6% and 1.8% higher than the baseline methods. Thus, it demonstrates robust regression and classification performance.

1. Introduction

With the widespread adoption of mobile devices and increasing pervasiveness of social networks, social media platforms have emerged as primary channels for individuals to express thoughts and opinions. These platforms host vast quantities of personal, emotionally charged data encompassing text, voice, image, and video. Comprehensively mining and analyzing users' "emotional data" across these multimodal sources not only supports refined emotion-driven marketing and personalized recommendations but also facilitates services tailored to customers' needs. Moreover, such analysis strengthens public-opinion monitoring and can serve as a powerful reference for government decision-making [1].

Traditionally, studies have primarily focused on analyzing emotions within single data modalities [2,3], which may not offer sufficient information in terms of emotional nuances. For instance, relying solely on text data might obscure crucial cues such as voice tone, facial expressions, or body language, all of which are integral to a comprehensive understanding of emotional states. There is growing research interest in sentiment analysis based on the fusion of multimodal data, given the inherent complementarity between modals [4].

In multimodal sentiment analysis research, features are typically extracted across multiple modals to acquire representative information from each modality. Subsequently, suitable algorithms can be employed to integrate these diverse data, thereby generating a multimodal representation [5]. Conventional approaches often involve concatenating or stacking feature vectors from different modalities to form a more intricate vector or matrix, which is then fed into deep neural networks for training. While effective, these methods tend to impose heavy computational burdens and overlook the unique contributions of each modality to emotional analysis tasks, resulting in suboptimal performance [6].

Early efforts to resolve these problems (e.g., CATF-LSTM [7]) centered on constructing Long Short-Term Memory (LSTM) structures and attention networks to simulate contextual information in various modes. Recent studies leveraging the Transformer model, such as CTFN [8] and MulT [9], have employed the concept of modality transformation via Transformer-based structures to facilitate intermodal encoding.

Approaches rooted in representation learning, such as MISA [10] and Self-MM [11], consider both the consistency and divergence of information across modes simultaneously. This enhances the accuracy of multimodal sentiment analysis. Despite these advancements, most existing methods disregard the impact of redundant features, treat all modalities uniformly, and lack a comprehensive consideration of

* Corresponding author.
E-mail addresses: liuzijun@mail.ynu.edu.cn (Z. Liu), caili@ynu.edu.cn (L. Cai), ywj1322@ynu.edu.cn (W. Yang), hanks@ynu.edu.cn (J. Liu).

https://doi.org/10.1016/j.patcog.2024.110847
Received 4 April 2024; Received in revised form 23 July 2024; Accepted 25 July 2024
Available online 27 July 2024

feature redundancy within multimodal datasets, as well as the distinct contributions of individual modalities. To address these issues, a sentiment analysis method rooted in text information enhancement and multimodal feature fusion (TIEMFF) was developed in the present study. The primary contributions of this work can be summarized as follows.

(1) A multimodal sentiment feature extraction method was established based on the K-means clustering algorithm [12]. This method aims to identify similar features within modality data and merge them to minimize redundancy. This enhances both the efficiency and accuracy of the model, while aiding in revealing commonalities and differences in emotional expressions across modalities.

(2) The proposed multimodal sentiment analysis method, based on attention mechanisms and cross-modal interactions, incorporates a fusion algorithm for cross-modal interactions and an attention mechanism during the fusion process. By alleviating interference from redundant and noisy information during modality interactions, this method enhances emotional interactions between modalities to improve the effectiveness of the recognition algorithm.

The rest of this paper is organized as follows. Section 2 describes related work and summarizes existing research results in multimodal sentiment analysis. Section 3 presents the details of the single-text modal sentiment analysis module. Section 4 details the proposed multimodal sentiment feature extraction and fusion method. Section 5 presents the experimental procedure, experimental results, and analyses. Finally, Section 6 concludes and suggests directions for further research.

2. Related work

The framework of multimodal sentiment analysis can be divided into intra- and inter-modal modeling [13]. Intra-modal modeling involves analyzing interactions within each modality, while inter-modal modeling deals with interactions between different modalities. Inter-modal modeling is also categorized as synchronous or asynchronous, depending on whether the modalities are aligned [14]. When modalities are already aligned, synchronous modeling methods can be employed to analyze information from different modalities at the same timestamp. If modalities lack alignment, asynchronous modeling methods are necessary to handle the occurrences of different modalities at different times. For example, alignment tools (such as P2FA [15]) can be used to align modalities before modeling.

2.1. Intra-modal modeling methods

Currently, a common approach for multimodal sentiment analysis is to utilize feature-based multimodal fusion methods for intra-modal modeling. These methods involve integrating features from different modalities at a shallow level after feature extraction. This fusion process entails mapping features from different unimodal sources into a shared parameter space. Due to potential variations in the information content across modalities, the resulting features often contain redundant information. To address this problem, dimensionality reduction techniques are often used to reduce redundancy, and the simplified features are fed into the model for subsequent feature extraction and prediction. However, feature-based multimodal fusion tends to lead to inadequate modeling of the dynamics within each modality and lacks comprehensive consideration of the details therein, thus losing the contextual and temporal dependencies within each modality, which in turn affects the modeling of cross-modal dynamics and leads to data overfitting. Majumder et al. [16] presented a novel feature strategy that proceeded in a hierarchical fashion, first fusing the modalities two in two and only then fusing all three modalities. Subsequently, leveraging fully connected layers and recurrent neural network (RNN) for context propagation facilitated effective sentiment classification tasks. However, this method treated each modality equally and failed to focus on modalities that contribute significantly, such as text modality. In tackling redundant information in sentiment analysis, Li et al. [17] introduced a model based on cross-modal semantic information enhancement. They achieved multimodal feature fusion through BiLSTM network and cross-modal information interaction mechanism. The utilization of multi-head self-attention mechanism enhanced the recognition of crucial sentiment features while reducing interference from redundant information, but they did not consider integrating sentiment knowledge from texts into multimodal systems.

2.2. Inter-modal modeling methods

For inter-modal modeling, a popular approach is to employ decision-level fusion-based multimodal fusion methods. Decision-level fusion involves training independent models using data from different modalities and combining their outputs to make the final decision. Averaging, majority voting, weighted fusion, and learnable models are usually used to integrate information from different modalities. In cases where certain modalities are unavailable, decisions can still be made using other available modalities. However, due to the construction of separate models to handle each modality, it is often challenging to effectively model the interactions between modalities. Luo et al. [18] proposed an adaptive weight fusion method for multimodal sentiment analysis. This method addressed issues arising from imbalanced modality learning rates, which could negatively impact the collaborative decision-making in multimodal analysis. Although this method improved the accuracy of sentiment analysis by balancing the learning rates of different modalities, it did not fully explored the complex interaction relationships between modalities. Zhang et al. [19] proposed a data-driven multimodal sentiment analysis framework based on a weight-based feature fusion method to capture more modality-specific information. Noisy modalities or those that contributed little to multimodal sentiment analysis decisions were given less fusion weight while preserving modality-specific information. However, it enhanced the computational sources and high training time. Wang et al. [20] proposed a lightweight Loss Switching Fusion Network (LSFNet) that dynamically selected appropriate loss functions for feature fusion based on the evolving conditions during the training process. By switching between different loss functions during training, the model can be guided to better learn the relationships between different modalities' features.

2.3. Multimodal fusion methods

Multimodal fusion methods typically include multi-kernel learning, neural networks, and graph models. Mai et al. [21] introduced a three-modal representation learning framework based on contrastive learning. They comprehensively learned intra- and inter-modal modeling through intra-modal, inter-modal, and semi-contrastive learning approaches. The incorporation of refinement terms and modality residuals enhanced the capability of representation learning. However, this study focused on contrastive learning between unimodal representations and had not yet innovated multimodal fusion methods, resulting in limited explanatory power of multimodal fusion. Ye et al. [22] presented a sentiment-aware multimodal pre-training (SMP) framework for sentiment analysis. This framework incorporated a cross-modal contrastive learning module to capture interactions between visual and textual information. However, when processing image-text pairs, the SMP model only utilized one adjective-noun pair per pair, thereby failing to fully leverage all possible descriptions, which may result in the omission of important sentiment information. Sun et al. [23] proposed a model based on cross-modal joint-encoding for capturing essential features between two modalities. However, extracting unimodal feature solely through modal alignment and one-dimensional convolution operation have not considered the impact of redundant features on multimodal sentiment analysis. He et al. [24] implemented the DISRFN

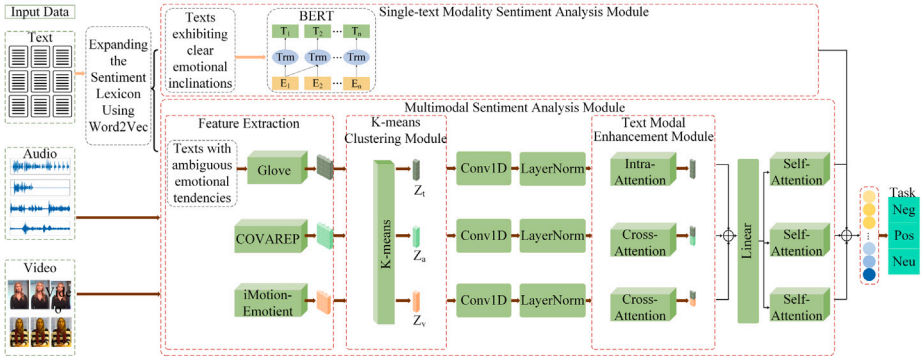

Fig. 1. TIEMFF architecture, composed primarily of "single-text modal sentiment analysis module", "multimodal feature extraction module", "modal feature clustering module based on *K*-means", and "text modal enhancement module (TE)" for "positive", "negative", and "neutral" classification tasks.

model, which leveraged the joint domain separation representations of all modes were obtained through the improved joint domain separation network. Then, the hierarchical graph fusion net (HGFN) was used for dynamically fusing each representation to obtain the interaction of multimodal data. This approach addressed the challenge of capturing the interactions between heterogeneous data sources, but it did not pay attention to the problem of multimodal data imbalance.

In summary, existing studies still encounter several challenges and limitations: such as how to effectively characterize the internal features of each modality while eliminating unnecessary redundant information and noise interference, how to fuse information between different modalities, and how to discover the feature differences between different modalities and highlight these differences. In this study, we propose a novel TIEMFF model, which can effectively eliminate redundant information and noise interference. In addition, by introducing text information enhancement, rich semantic features can be extracted from the text data. Finally, by utilizing cross modal interaction and attention mechanism, TIEMFF can capture the interactions and common features between different modalities, and deeply explore the diversity of sentiment expression.

3. Methods

The overall framework of the proposed TIEMFF model is illustrated in Fig. 1. TIEMFF primarily comprises two components. The first is a single-text modality sentiment analysis module, which individually trains the text modality with evident emotional inclinations using sentiment lexicons and the pre-trained bidirectional encoder representations from transformers (BERT) model. The second is a multimodal feature extraction module, which performs data processing and modality enhancement. Original feature sequences are extracted from text, audio, and video data using GloVe [25], COVAREP [26], iMotion-Emotient [9] methods, respectively, to obtain multimodal feature representations. The *K*-means clustering algorithm [12] is employed to reduce data redundancy by reconstructing highly similar modality data. The features are categorized into video-text and acoustic-text pairs, incorporating cross-attention and intra-attention mechanisms for the text modality. Finally, the pairs are fused with an adaptive fusion mechanism.

3.1. Single-text modality sentiment analysis module

Compared to audio and video modalities, the text modality most significantly influences multimodal sentiment analysis tasks [27,28]. Hence, the text modality was trained separately in this study, focusing on instances with distinct emotional inclinations within the dataset. Most studies utilize SentiWordNet [29], an ontology of sentiment words developed by Princeton University, as the foundation for sentiment analysis. SentiWordNet associates each word with sentiment polarity, intensity, and related emotional themes. Each word is categorized into positive, negative, or neutral sentiment classes with corresponding sentiment intensity scores. To optimize the sentiment vocabulary coverage and improve the adaptability of sentiment expressions to specific domains or contexts, thereby enhancing the sentiment analysis model's performance, the SentiWordNet lexicon was extended to generate the dataset used in this study. To operate this extension process:

(1) Tokenize the text within the dataset and incorporate sentiment words into the tokenizer's user dictionary to ensure the accurate recognition and retention of these terms;

(2) Construct a word set from the tokenized text data, arranging them in reverse order based on word frequency, and extract higher-frequency word sets as candidates;

(3) Manually review the data to identify new sentiment words from the candidate word sets;

(4) Incorporate the selected new sentiment words into the user dictionary to ensure their proper recognition in subsequent processing steps;

(5) Utilize the Word2vec model for word vector training on the constructed word sets, and expand these sets using the appearance of sentiment words in the training text;

(6) Manually review the expanded word sets to derive the final sentiment word set.

When processing the dataset, based on the final sentiment word set, the text data was categorized into two groups: texts exhibiting clear emotional inclinations and texts with ambiguous emotional tendencies. In the former group, as depicted in Fig. 2, texts were included if the PosScore or NegScore in their sentiment word set was $\geqslant 0.5$. In the latter group, as shown in Figure 3, texts were included if the PosScore or NegScore in their sentiment word set was < 0.5. For texts demonstrating clear emotional tendencies, a unimodal data sentiment recognition algorithm based on sentiment lexicons and the BERT model was applied.

The pre-trained BERT model [30] extracted sentence features and optimized the model by minimizing the cross-entropy loss function. This ensured the output probability distribution of the model was as close as possible to the actual labels, thereby enhancing the model's accuracy.

4. Multimodal sentiment feature extraction and fusion methods

4.1. Notations

In the dataset, each video clip containing a collection of video frames is assigned an overall sentiment label. Accordingly, in this study, a model was constructed to analyze the sentiment information using the text, audio, and video signals in the video clips. Features from different modalities were extracted from each video clip as model inputs. The corresponding notations are listed in Table 1.

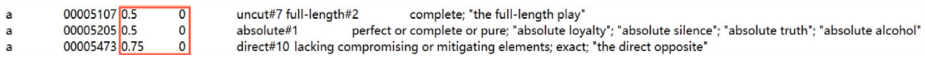

Fig. 2. Words with more pronounced affective tendencies in SentiWordNet lexical ontology, i.e., PosScore or NegScore $\geqslant 0.5$.

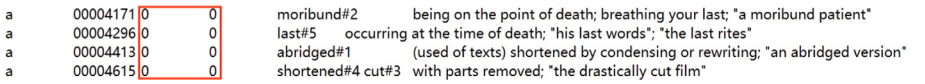

Fig. 3. Words with insignificant affective tendencies in SentiWordNet lexical ontology, i.e., PosScore or NegScore < 0.5.

Table 1
Notations for feature extraction and fusions.

Notation	Description
X	Multimodal datasets.
t, a, v	Text, audio, and video.
$X_t \in R^{T_t \times d_t}, X_a \in R^{T_a \times d_a}, X_v \in R^{T_v \times d_v}$	Text data, audio data, and video data.
T_t, T_a, T_v	Length of text sequence, length of audio sequence, and length of video sequence.
d_t, d_a, d_v	Dimensionality of word vectors for text generation, number of acoustic features extracted at each time point, and number of facial emotion features extracted at each time point.
F_t, F_a, F_v	Text features expressed as word vectors extracted from X_t, audio features expressed as time series extracted from X_a, and video features expressed as time series extracted from X_v.
Z_t, Z_a, Z_v	Clustering results for F_t, F_a, and F_v.
H_t, H_a, H_v	Representation of text, audio and video modality features after Conv1D.
G_t	Weight vector.

4.2. Multimodal feature extraction

GloVe [25] was utilized for feature extraction from the text modality. GloVe is a word-embedding method widely used in natural language processing. It operates by constructing a vocabulary, calculating co-occurrence statistics, training word vectors, and mapping these vectors into a lower-dimensional vector space. This process embeds vocabulary words from textual data into a lower-dimensional vector space, revealing semantic relationships among words. It is formulated as follows:

$$F_t = Glove\left(X_t; \theta_t^{Glove}\right) \in R^{T_t \times d_t}, \tag{1}$$

where θ_t^{Glove} represents the parameters of the GloVe method.

The COVAREP [26] toolkit was applied to extract audio features from the audio modality, covering fundamental frequency (F0), resonance frequencies, voice intensity, spectral envelope features, and other acoustic characteristics. Extracted features are represented in the form of a time series, where each time point corresponds to a set of acoustic features.

$$F_a = COVAREP\left(X_a; \theta_a^{COVAREP}\right) \in R^{T_a \times d_a}, \tag{2}$$

where $\theta_a^{COVAREP}$ represents the parameters used for extraction.

iMotion-Emotient [9] was applied to extract features from the video modality. iMotion-Emotient is a tool used for analyzing facial expressions in videos, extracting features in a series of steps such as face detection, facial expression analysis, eye feature analysis, and mouth feature analysis. Extracted facial features are represented in the form of a time series, where each time point corresponds to a set of emotion features. This relationship is expressed in Eq. (3):

$$F_v = iMotion - Emotient\left(X_v; \theta_v^{iMotion-Emotient}\right) \in R^{T_v \times d_v}, \tag{3}$$

where $\theta_v^{iMotion-Emotient}$ represents the parameters used for feature extraction. The feature extraction map for each modality is shown in Fig. 4.

4.3. Modality feature clustering based on K -means

The K-means clustering algorithm [12] was introduced to reduce data dimensionality, capturing similar features within the multimodal data and consolidating them to minimize redundancy among features and optimize the subsequent representation and analysis. The following equations represent the clustering of extracted modality features:

$$Z_t = Kmeans\left(F_t, K\right), \tag{4}$$

$$Z_a = Kmeans\left(F_a, K\right), \tag{5}$$

$$Z_v = Kmeans\left(F_v, K\right), \tag{6}$$

Following the clustering process, the generated Z_t, Z_a and Z_v contain reduced-dimensional representations of text, audio, and video features. This reduction in dimensionality compresses the original features into fewer feature vectors, capturing the information present in the original data. The reduced-dimensional data can be employed for sentiment analysis tasks, where each sample is represented by a label belonging to a specific cluster. This approach retains the essential features of the data while minimizing redundancy, thereby enhancing the accuracy and computational efficiency of the sentiment analysis process.

4.4. Multimodal sentiment analysis module

In the proposed method, texts with ambiguous emotional tendencies are passed through a multimodal sentiment feature extraction module based on the K-means clustering algorithm to process them and acquire an input sequence. This input sequence is then passed through a one-dimensional (1D) temporal convolutional layer to ensure each element comprehensively perceives information from its neighboring elements and standardizes dimensions.

$$H_t = Conv1D\left(Z_t, kernel_t\right), \tag{7}$$

$$H_a = Conv1D\left(Z_a, kernel_a\right), \tag{8}$$

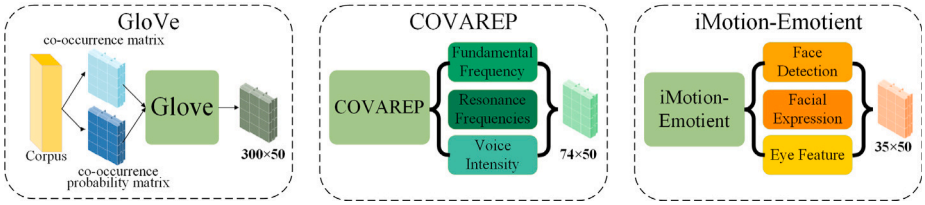

Fig. 4. Feature extraction methods of text modality, audio modality, and video modality.

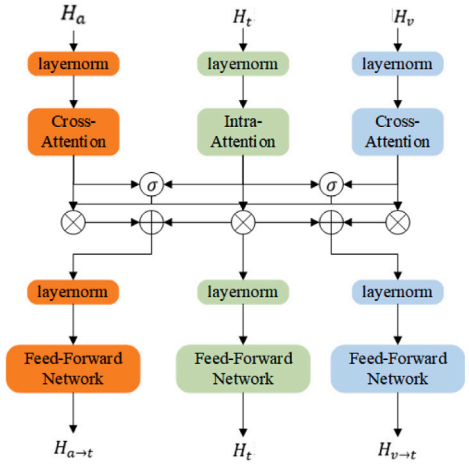

Fig. 5. TE module architecture consisting of "cross-attention mechanism" and "intra-attention mechanism" for cross-modal fusion.

$$H_v = Conv1D\left(Z_v, kernel_v\right), \tag{9}$$

where H_t, H_a and H_v are integrated representations of text, audio, and video modalities, respectively; $kernel_t$, $kernel_a$ and $kernel_a$ denote the kernel sizes of the 1D convolutional networks for each modality. Data from different modalities is processed and fused through the 1D convolutional layer, ensuring that each element thoroughly perceives information from its adjacent elements, allowing features from various modalities to be expressed in a unified feature space. This unified representation of features ensures consistency subsequent sentiment analysis tasks without sacrificing comprehensiveness.

The text modal enhancement module (TE), as illustrated in Fig. 5, consists of two cross-attentions and an intra-attention mechanism. Initially, the features H_t, H_a and H_v undergo normalization processing. Subsequently, these normalized features are fed into the attention modules. Cross-attention takes H_a and H_v as inputs, while intra-attention utilizes H_t as an input. Through an adaptive fusion mechanism, the enhanced features are processed to produce a weighted fusion of feature representations. This approach thoroughly integrates emotional information from multimodal data, effectively fusing different modalities.

Eq. (10) describes the cross-attention mechanism, which employs adaptive fusion for feature enhancement. Notably, the audio and video modalities share the same text-based intra-attention block, so H_t is integrated into the audio and video modality features separately; this generates enhanced representations denoted as $H_{a\to t}$ and $H_{v\to t}$. Here, σ represents the sigmoid function, while L_t, $L_{a\to t}$ and $L_{v\to t}$ denote learnable parameters. The model utilizes these learnable parameters to determine the weights for each corresponding feature dimension of H_t and $H_{a\to t}$, $H_{v\to t}$. These weights are used to attain G_i for the weighted fusion of H_t and $H_{a\to t}$, $H_{v\to t}$.

$$G_i = \sigma\left(H_t \times L_t + H_{a\to t} \times L_{a\to t} + + H_{v\to t} \times L_{v\to t}\right), \tag{10}$$

Eq. (11) illustrates the intra-attention mechanism. Within this mechanism, the weighted-fused enhanced features $H_{a\to t}$, $H_{v\to t}$ and H_t are utilized. The variable $\odot$ signifies element-wise multiplication. Through element-wise multiplication, G_i is multiplied with H_t and $H_{a\to t}$, $H_{v\to t}$ for a weighted fusion of features. This operation determines the contribution of text-based features and enhanced audio-visual features on each feature dimension based on G_i. If the weight of G_i is close to 1, the contribution of the H_t is relatively high, while the contribution of the $H_{a\to t}$, $H_{v\to t}$ is relatively low. Conversely, if the weight of G_i is close to 0, the contribution of the $H_{a\to t}$, $H_{v\to t}$ is relatively high, and the contribution of the H_t is relatively low. This mechanism effectively regulates the balance among text and enhanced features across different dimensions during the feature fusion process, catering to the requirements for diverse emotional expressions.

$$H_{a\to t} = G_i \odot H_t + \left(1 - G_i\right) \odot H_{a\to t}, \tag{11}$$

$$H_{v\to t} = G_i \odot H_t + \left(1 - G_i\right) \odot H_{v\to t}, \tag{12}$$

Fig. 6 shows the details of cross-modal fusion using text modality and audio modality as examples. First, for text features, the intra-attention mechanism is applied to extract key information. Next, H_a is fused with H_t to generate the enhanced feature $H_{a\to t}$. In this step, attention scores are computed and normalized to obtain weights. These weights are used to weight the fused augmented feature $H_{a\to t}$ for further fusion with H_t. The fused features are dynamically regulated by the proportion of weights to determine the importance of text features and augmented features. Finally, the fused output is processed with the softmax function.

$$H_{a\to t} = FFN\left(LN\left(H_{a\to t}\right)\right), \tag{13}$$

$$H_{v\to t} = FFN\left(LN\left(H_{v\to t}\right)\right), \tag{14}$$

$$H_t = FFN\left(LN\left(H_t\right)\right), \tag{15}$$

Eqs. (13) to (15) demonstrate the operation of the feedforward neural network (FFN). $H_{a\to t}$, $H_{v\to t}$ represent the enhanced features processed through the FFN, while H_t signifies the text-based features processed through the same network. Layer normalization (LN) normalizes the input features. The enhanced features $H_{a\to t}$ and $H_{v\to t}$, alongside text-based features H_t, are processed through the FFN, enabling a deeper integration of multimodal data and comprehensive learning of emotional information.

The unified-dimensional features of the three modalities are normalized to enhance the model's robustness and ensure suitable gradients. Normalization confines the feature values within an appropriate range, stabilizing the model and improving training efficacy. Information within each modality complements cross-modal information, so the model was equipped with a self-attention layer to sequentially gather

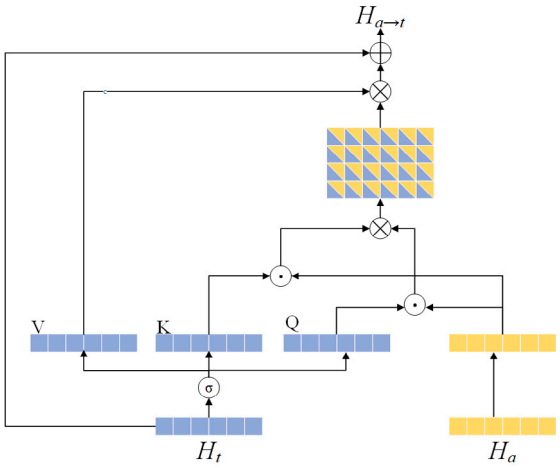

Fig. 6. Cross-modal fusion module architecture. Taking audio modality as an example, blue represents text features, yellow represents audio features, and Q, K, and V represent the query matrix, key matrix, and value matrix of the self-attention mechanism, respectively, illustrating details of cross-modal fusion.

Table 2
Dataset statistics in CMU-MOSI and CMU-MOSEI.

Dataset	Data category	Data items
CMU-MOSI	Train	1283
	Valid	299
	Test	686
CMU-MOSEI	Train	16 326
	Valid	1871
	Test	4659

temporal information. This self-attention layer interlinks the reinforced features of the three modalities, then the last element is extracted from the sequence model as the feature representation. Finally, a fully connected layer can be applied to the extracted features, completing the sentiment analysis task.

5. Experiment

5.1. Experimental environment and dataset

The experimental setup for this study included a DELL Precision 5820 Tower X-Series equipped with an Intel (R) Core (TM) i7-9800X CPU @ 3.80 GHz processor, 64 GB of memory, and an NVIDIA GeForce RTX 2080 Ti graphics card with 11 GB of memory. Windows 10 was used as the operating system. The work was conducted in Python 3.7 programming language and the PyTorch framework was employed to train the parameters of deep learning networks within the Windows environment.

To validate the experimental results, two publicly available datasets for multimodal sentiment analysis, CMU-MOSI [31] and CMU-MOSEI [32], were used as benchmarks. CMU-MOSI, provided by Carnegie Mellon University, is a commonly used dataset in multimodal sentiment analysis research. To facilitate its operation, annotators manually assigned emotion labels ranging from −3 to 3 to each segment sample in the dataset. CMU-MOSEI, an upgraded version of CMU-MOSI, is currently the largest-scale multimodal sentiment analysis dataset available; it comprises 353,2928 segmented video clips extracted from 23 videos, including emotional intensity labels and multiple emotion categories (e.g., “happiness”, “sadness”). The partitioning details for both datasets are presented in Table 2.

Table 3
Experimental parameters.

Hyperparameter	Value
K	3
$n$levels	5
batch_size	24
Learning_rate	1e−5
num_epochs	43
dropout	0.1
Optimization Function	Adam
Loss Function	L1Loss

5.2. Evaluation metrics

Two evaluation tasks – regression and classification – were performed in these experiments. The results of the regression task are presented here in terms of the mean absolute error (MAE). For the classification task, multiple evaluation metrics such as F1 score, binary classification accuracy, and Pearson correlation coefficient (Corr) were applied. Higher values indicate better performance across all metrics except for MAE, where lower values signify better performance.

5.3. Experimental parameter setting

Table 3 provides a detailed list of experimental parameters, including the number of clusters (K), network layers, batch size, learning rate, epochs, dropout rate, optimization function, and loss function. The learning rate markedly impacts the step size of parameter updates during gradient descent, affecting the speed of the loss function descent as well as model convergence. As shown in Fig. 7, when Learning_rate is 1e-5, the learning rate curve demonstrates a gradual descent trend towards convergence around Epoch=43. This learning rate trend suggests that the model did not encounter issues of exploding or vanishing gradients.

The choice of K value significantly influences the K-means clustering algorithm’s performance. The Elbow Method [33] was utilized in this study to determine the most suitable K value. This involved plotting the relationship between the loss function (also known as the “within-cluster sum of squares”) and varying K values, aiming to identify an “elbow” point reflective of the optimal K value. In experiments conducted on the CMU-MOSEI dataset, this elbow in the curve emerged at $K = 3$ (Fig. 8) prior to the onset of a smoother trend. Based on this turning point, the optimal K value is 3. Similarly, in the experiments conducted on the CMU-MOSI dataset, an elbow in the curve was observed at $K = 3$ (Fig. 9), again indicating an optimal K value of 3.

5.4. Experimental results and analysis

5.4.1. Benchmark method

To validate the effectiveness of the proposed TIEMFF method, comparisons with various benchmark methods were conducted as described in Table 4.

5.4.2. Algorithm comparison

Four metrics were evaluated in the comparative experiments: binary classification accuracy, Corr, F1 score, and MAE. Fig. 10 illustrates the experimental results obtained for the CMU-MOSI dataset and Fig. 11 presents those for the CMU-MOSEI dataset.

Based on the experimental outcomes depicted in Figs. 10 and 11, the TIEMFF method demonstrates significant performance advantages in sentiment classification tasks on both CMU-MOSI and CMU-MOSEI datasets. Its comprehensive modality fusion framework lends superior performance, integrating single-text modality sentiment analysis with multimodal data processing. However, on the CMU-MOSI

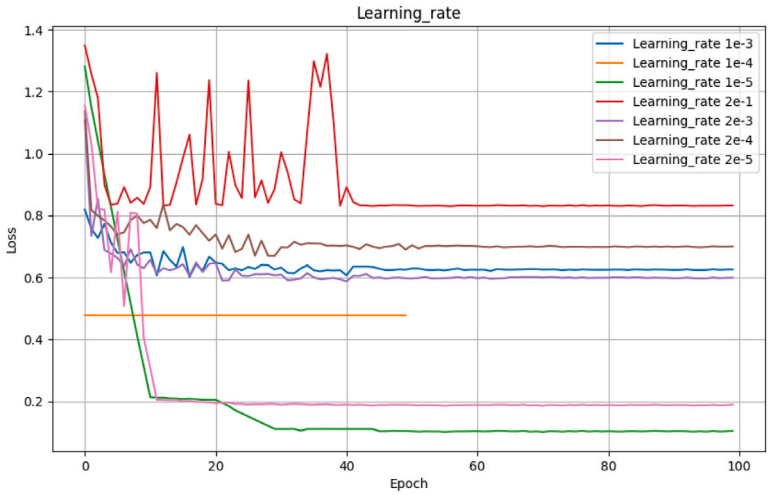

Fig. 7. Learning_rate values and loss change curves. The blue curve represents the variation of loss with a Learning_rate of 10^{-3}; the yellow curve represents the variation of loss with a Learning_rate of 10^{-4}; the green curve represents the variation of loss with a Learning_rate of 10^{-5}; the red curve represents the variation of loss with a Learning_rate of 2×10^{-1}; the purple curve represents the variation of loss with a Learning_rate of 2×10^{-3}; the brown curve represents the variation of loss with a Learning_rate of 2×10^{-4}; and the pink curve represents the variation of loss with a Learning_rate of 2×10^{-5}.

Table 4
Introduction to baseline methodologies.

Benchmark method names	Benchmark method description
EF-LSTM	Benchmark method dedicated to entity-level relationship extraction, focuses attention on entities in a sentence and utilizes LSTM networks to capture semantic information and sequence structures.
LF-LSTM	Benchmark method for entity-level relationship extraction, focuses on local features in sentences and incorporates LSTM networks to capture semantic information and sequence structures.
MulT(2019) [9]	Multimodal transformer network, directly learns representation of unaligned multimodal data streams with a cross-attention module to directly process cross-modal associations of original features without manual word alignment preprocessing.
MAG-BERT(2020) [34]	Multimodal adaptive gate (MAG) as a BERT and XLNet accessory to accept multimodal nonverbal data during fine-tuning and incorporate multimodal information into their internal representations.
Self-MM(2021) [11]	Self-supervised learning strategy generates unimodal supervision and jointly trains multimodal and unimodal tasks to learn feature consistency and variability.
MMLatch(2022) [35]	MMLatch model employs an end-to-end framework that utilizes representations from higher-level layers within neural networks to create top-down masks for low-level input features; these masks are integrated into an architecture based on LSTM encoders and cross-attention, with an adaptive gating mechanism to integrate and regulate multimodal information.
COGMEN(2022) [36]	Utilizes a structure based on graph neural networks to model complex dependencies in dialogues, encompassing both local and global information; introduces contextual information to enhance sentiment recognition performance.
CHAMPAGNE-BASE(2023) [37]	CHAMPAGNE-BASE integrates a hybrid approach of graph neural networks and generative models, showcasing superior performance in open-domain textual conversations, understanding social interactions, and sentiment analysis.
GEAR(2023) [38]	Employs robust feature extraction, biased feature separation, and enhanced generalization to decouple multimodal features and estimate biased weights, thereby enhancing sentiment analysis performance.
TETFN(2023) [4]	Combines LSTM and temporal convolutional networks to encode contextual information for each modality, with a cross-modal transformer focused on modeling paired cross-modal mappings for each modality to generate unimodal labels, thus capturing unique differential information among modalities.

dataset, the TIEMFF method exhibits slightly higher MAE metrics compared to other models, indicating a need for further enhancements in certain sentiment classification scenarios. Conversely, on the CMU-MOSEI dataset, the TIEMFF method displays a 5.4% reduction in F1 score compared to the CHAMPAGNE-BASE model, indicating notable performance differences on specific sentiment classification tasks.

5.5. Ablation experiment

A series of comparative experiments were conducted to explore the effects of different modules within the TIEMFF model, as delineated in Tables 5 and 6. Among the three singular modalities, accuracy for the video modality was slightly higher than for the audio modality. This is likely due to the richer information encapsulated

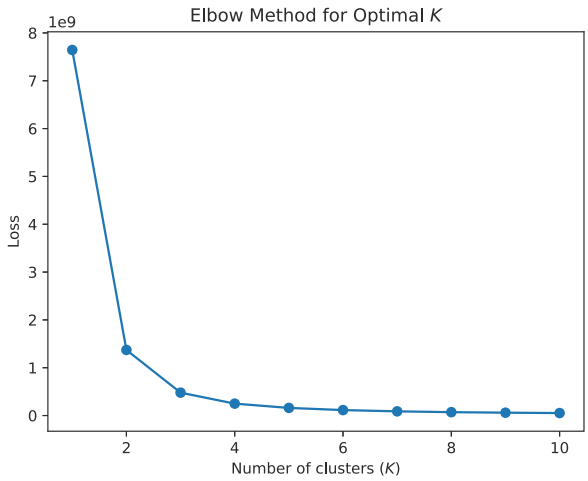

Fig. 8. K value change diagram of CMU-MOSEI data set based on Elbow Method.

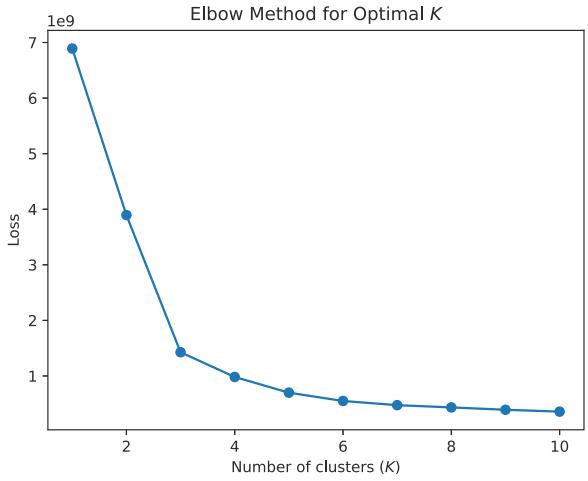

Fig. 9. K value change diagram of CMU-MOSI data set based on Elbow Method.

Table 5
Ablation experiments conducted on CMU-MOSI dataset.

Model	Acc2	F1 score	MAE	Corr
Only T	79.1	78.2	0.853	0.531
Only A	65.6	68.8	0.864	0.310
Only V	66.4	69.3	0.859	0.343
TE	83.7	82.1	0.747	0.656
TE + SentiWordNet + Bert	84.3	83.5	0.739	0.674
TE + SentiWordNet + Bert + K-Means	**85.7**	**85.8**	**0.727**	**0.698**

in video data, encompassing facial expressions, body language, and environmental information, providing more cues for sentiment analysis. However, the performance on the single-text modality was far superior to both video and audio modalities, further supporting the dominance of text modality in sentiment analysis tasks. The multimodal text modal enhancement module facilitated the integration of complementary information between different modalities, enhancing the expression of sentiment in videos and improving the accuracy of the analysis. Subsequent integration of the single-text sentiment analysis module further enhanced accuracy, indicating that multimodal sentiment recognition can effectively utilize information from different modalities to comprehensively capture emotional content. Furthermore, the K-means clustering algorithm maintains sample diversity while minimizing reduced data redundancy, streamlining sentiment analysis tasks.

Table 6
Ablation experiments conducted on CMU-MOSEI dataset.

Model	Acc2	F1 score	MAE	Corr
Only T	62.5	72.0	0.724	**71.2**
Only A	57.8	76.5	0.898	64.6
Only V	68.1	79.7	0.844	62.8
TE	82.5	79.5	0.682	69
TE + SentiWordNet + Bert	84.3	78.5	0.651	70.2
TE + SentiWordNet + Bert + K-Means	**85.9**	**81**	**0.548**	70.5

A further comparison was conducted to explore the significance of the multimodal text modality enhancement module by evaluating different mappings, including single cross-modal mappings ($[a,v] \rightarrow t$, $[a,t] \rightarrow v$, $[v,t] \rightarrow a$), bi-cross-modal mappings ($[v,t] \rightarrow a$ and $[a,v] \rightarrow t$, $[v,t] \rightarrow a$ and $[a,t] \rightarrow v$, $[a,v] \rightarrow t$ and $[a,t] \rightarrow v$), and tri-modal inter-mappings $[at, vt, av]$. The results, as presented in Tables 7 and 8, indicate distinct performance among these mappings. In the single cross-modal mappings, $[a,t] \rightarrow v$ and $[v,t] \rightarrow a$ performed worse compared to $[a,v] \rightarrow t$. The tri-modal inter-mapping exhibited higher accuracy than the bi-cross-modal mappings but slightly lower accuracy than $[a,v] \rightarrow t$. This suggests that text contributes more significantly to multimodal sentiment analysis. Compared to audio and video modalities, text carries greater value in expressing emotional information to support multimodal sentiment analysis.

6. Conclusion

This paper presents a novel approach to multimodal sentiment analysis, aimed at adapting to varying demands by reducing redundancy and adjusting the trade-offs and contributions of text and enhanced features across different dimensions in the feature fusion process. An expanded SentiWordNet lexical ontology was first employed to differentiate between texts with explicit emotional tendencies and those with only indirect (or ambiguous) emotional expressions. For texts with clear emotional tendencies, a single-text modality sentiment analysis module was utilized, integrating the SentiWordNet lexical ontology and pre-trained BERT model to analyze the emotional inclinations. This approach allows the model to focus on the emotional expression within these texts, avoiding interference from other modality data and enabling a more refined capture and analysis of emotional features. For texts with indirect emotional expressions, a multimodal sentiment analysis was employed, leveraging GloVe, COVAREP, and iMotion-Emotient techniques to extract raw feature sequences from the text, audio, and video modalities, respectively. Subsequently, the K-means clustering algorithm was applied to reduce redundancy in the multimodal data, which enhances efficiency and reveals crucial features. A text modality enhancement module was then deployed to facilitate the integration and transfer of information within and across modalities, based on inter- and intra-attention mechanisms, to better capture correlations within the multimodal data.

Although the proposed model represents a significant advancement in multimodal sentiment analysis, it has persistent limitations.

Firstly, the current sentiment analysis mainly focuses on traditional modality information including text, audio and video, without fully covering other potentially valuable modality information. Therefore, future research should expand the modality scope of sentiment analysis by introducing more diverse modality information (including but not limited to physiological signals, images, and other perceptual modalities) to achieve a more comprehensive understanding of the diversity and complexity of emotional expressions.

Secondly, although the application of the proposed emotion lexicon was found to effectively improves the expression and understanding of emotion information, there are still some issues with its construction and usage. Future research may optimize the construction and application of the emotion lexicon to enhance its adaptability in different

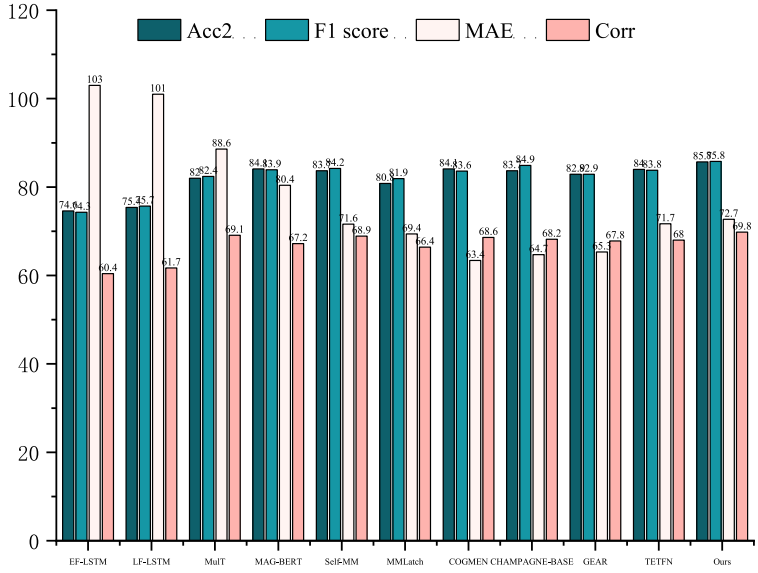

Fig. 10. TIEMFF algorithmic comparison results on CMU-MOSI dataset across evaluation metrics Acc2, F1 Score, MAE and Corr.

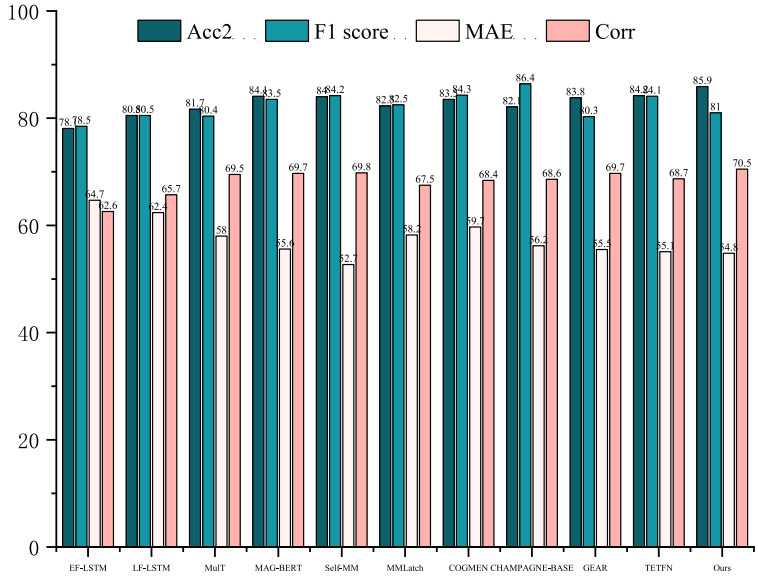

Fig. 11. TIEMFF algorithmic comparison on CMU-MOSEI dataset across evaluation metrics Acc2, F1 Score, MAE and Corr.

Table 7
Experimental results based on various quantities of modalities within CMU-MOSI dataset.

Number of modalities	Model	Acc2	F1 score	MAE	Corr
	$[a,v] \rightarrow t$	**85.7**	**85.8**	**0.727**	**69.8**
one single cross-modal	$[a,t] \rightarrow v$	83.9	82.9	0.772	67.6
	$[v,t] \rightarrow a$	83.6	82.5	0.743	68.1
	$[v,t] \rightarrow a$ and $[a,v] \rightarrow t$	84.6	84.3	0.735	68.4
bi-cross-modal	$[v,t] \rightarrow a$ and $[a,t] \rightarrow v$	84.2	82.9	0.740	67.8
	$[a,v] \rightarrow t$ and $[a,t] \rightarrow v$	84.5	82.7	0.756	61.3
tri-cross-modal	$[at, vt, av]$	85.1	83.8	0.871	68.8

contexts and fields. This can include making the emotion lexicon more adaptable to multilingual and multicultural environments, while considering customizing emotional vocabulary to improve accuracy and applicability in different fields and themes.

In addition, there is yet a need for more thorough empirical experiments to fully evaluate the proposed method's generalizability and stability in real-world scenarios. Future studies may include more comprehensive experiments, such as in-depth studies on testing on diverse datasets, considering the differences in sentiment expression across different scenarios and user groups, and evaluating the robustness of the model throughout long-term usage periods.

Funding

This work was supported in part by the National Natural Science Foundation of China under Grant No. 82360280, and the Open Foundation of Key Laboratory in Media Convergence of Yunnan Province under Grant No. 220225201.

Table 8
Experimental results based on various quantities of modalities within CMU-MOSEI dataset.

Number of modalities	Model	Acc2	F1 score	MAE	Corr
one single cross-modal	$[a, v] \rightarrow t$	**85.9**	**81**	**0.548**	**70.5**
	$[a, t] \rightarrow v$	68.1	71.9	0.791	32.2
	$[v, t] \rightarrow a$	70.4	71.9	0.785	36.6
bi-cross-modal	$[v, t] \rightarrow a$ and $[a, v] \rightarrow t$	81.2	81.3	0.648	66.7
	$[v, t] \rightarrow a$ and $[a, t] \rightarrow v$	74.0	73.8	0.765	45.3
	$[a, v] \rightarrow t$ and $[a, t] \rightarrow v$	80.5	80.6	0.645	64.9
tri-cross-modal	$[at, vt, av]$	68.7	71.6	0.799	31.6

CRediT authorship contribution statement

Zijun Liu: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Li Cai:** Writing – review & editing, Software, Methodology, Funding acquisition, Formal analysis. **Wenjie Yang:** Validation, Supervision, Methodology, Investigation. **Junhui Liu:** Supervision, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

No conflict of interest exits in the submission of this manuscript, and manuscript is approved by all authors for publication. I would like to declare on behalf of my co-authors that the work described was original research that has not been published previously, and not under consideration for publication elsewhere, in whole or in part. All the authors listed have approved the manuscript that is enclosed.

Data availability

The authors do not have permission to share data.

References

[1] Y. Ruyun, M. Jing, A feature-enhanced multi-modal emotion recognition model integrating knowledge and res-vit, Data Anal. Knowl. Discov. 7 (11) (2023) 14–25, http://dx.doi.org/10.11925/infotech.2096-3467.2022.1020.

[2] Q. Shi, J. Fan, Z. Wang, Z. Zhang, Multimodal channel-wise attention transformer inspired by multisensory integration mechanisms of the brain, Pattern Recognit. 130 (2022) 108837, http://dx.doi.org/10.1016/j.patcog.2022.108837.

[3] K. Ezzameli, H. Mahersia, Emotion recognition from unimodal to multimodal analysis: A review, Inf. Fusion (2023) 101847., http://dx.doi.org/10.1016/j.inffus.2023.101847.

[4] D. Wang, X. Guo, Y. Tian, J. Liu, L. He, X. Luo, TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis, Pattern Recognit. 136 (2023) 109259, http://dx.doi.org/10.1016/j.patcog.2022.109259.

[5] N. Xu, W. Mao, MultiSentiNet: A deep semantic network for multimodal sentiment analysis, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Association for Computing Machinery, 2017, pp. 2399–2402, http://dx.doi.org/10.1145/3132847.3133142.

[6] Y. Yin, L. Jing, F. Huang, G. Yang, Z. Wang, Msa-gcn: Multiscale adaptive graph convolution network for gait emotion recognition, Pattern Recognit. 147 (2024) 110117, http://dx.doi.org/10.1016/j.patcog.2023.110117.

[7] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, L.-P. Morency, Multi-level multiple attentions for contextual multimodal sentiment analysis, in: 2017 IEEE International Conference on Data Mining, ICDM, IEEE, 2017, pp. 1033–1038, http://dx.doi.org/10.1109/ICDM.2017.134.

[8] J. Tang, K. Li, X. Jin, A. Cichocki, Q. Zhao, W. Kong, CTFN: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 5301–5311, http://dx.doi.org/10.18653/v1/2021.acl-long.412.

[9] Y.-H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the Conference. Association for Computational Linguistics. Meeting, vol. 2019, NIH Public Access, 2019, p. 6558, http://dx.doi.org/10.18653/v1/p19-1656.

[10] D. Hazarika, R. Zimmermann, S. Poria, Misa: Modality-invariant and-specific representations for multimodal sentiment analysis, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1122–1131, http://dx.doi.org/10.1145/3394171.3413678.

[11] W. Yu, H. Xu, Z. Yuan, J. Wu, Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, (no. 12) 2021, pp. 10790–10797, http://dx.doi.org/10.1609/aaai.v35i12.17289.

[12] M. Ahmed, R. Seraj, S.M.S. Islam, The k-means algorithm: A comprehensive survey and performance evaluation, Electronics 9 (8) (2020) 1295, http://dx.doi.org/10.3390/electronics9081295.

[13] Z. Lin, B. Liang, Y. Long, Y. Dang, M. Yang, M. Zhang, R. Xu, Modeling intra- and inter-modal relations: Hierarchical graph contrastive learning for multimodal sentiment analysis, in: Proceedings of the 29th International Conference on Computational Linguistics, vol. 29, (no. 1) Association for Computational Linguistics, 2022, pp. 7124–7135.

[14] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, X. Kong, Multimodal sentiment analysis based on fusion methods: A survey, Inf. Fusion 95 (2023) 306–325, http://dx.doi.org/10.1016/j.inffus.2023.02.028.

[15] S.A. Abdu, A.H. Yousef, A. Salem, Multimodal video sentiment analysis using deep learning approaches, a survey, Inf. Fusion 76 (2021) 204–226, http://dx.doi.org/10.1016/j.inffus.2021.06.003.

[16] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, S. Poria, Multimodal sentiment analysis using hierarchical fusion with context modeling, Knowl.-Based Syst. 161 (2018) 124–133, http://dx.doi.org/10.1016/j.knosys.2018.07.041.

[17] L. Mengyun, Z. Jing, Z. Huanxiang, Z. Xiaolin, L. Luyao, Multimodal sentiment analysis based on cross-modal semantic information enhancement, J. Front. Comput. Sci. Technol. (2023) 1–13., http://dx.doi.org/10.3778/j.issn.1673-9418.2307045.

[18] L. Yuanyi, W. Rui, L. Jiafeng, T. Xianglong, Multimodal sentiment analysis based on adaptive weight fusion, J. Softw. (2023) 1–13., http://dx.doi.org/10.13328/j.cnki.jos.006998.

[19] J. Zhang, X. Wu, C. Huang, AdaMoW: Multimodal sentiment analysis based on adaptive modality-specific weight fusion network, IEEE Access 11 (2023) 48410–48420.

[20] L. Wang, D.Q. Huynh, M.R. Mansour, Loss switching fusion with similarity search for video classification, in: 2019 IEEE International Conference on Image Processing, ICIP, IEEE, 2019, pp. 974–978, http://dx.doi.org/10.1109/ICIP.2019.8803051.

[21] S. Mai, Y. Zeng, S. Zheng, H. Hu, Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis, IEEE Trans. Affect. Comput. 14 (3) (2022) 2276–2289, http://dx.doi.org/10.1109/TAFFC.2022.3172360.

[22] J. Ye, J. Zhou, J. Tian, R. Wang, J. Zhou, T. Gui, Q. Zhang, X. Huang, Sentiment-aware multimodal pre-training for multimodal sentiment analysis, Knowl.-Based Syst. 258 (2022) 110021, http://dx.doi.org/10.1016/j.knosys.2022.110021.

[23] S. Bin, J. Tao, J. Li, Multimodal sentiment analysis based on cross-modal joint-encoding, Comput. Eng. Appl. (2023) 1–10, http://dx.doi.org/10.3778/j.issn.1002-8331.2306-0364.

[24] J. He, H. Yanga, C. Zhang, H. Chen, Y. Xua, et al., Dynamic invariant-specific representation fusion network for multimodal sentiment analysis, Comput. Intell. Neurosci. 2022 (2022) http://dx.doi.org/10.1155/2022/2105593.

[25] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, Springer, 2014, pp. 1532–1543, http://dx.doi.org/10.3115/v1/d14-1162.

[26] G. Degottex, J. Kane, T. Drugman, T. Raitio, S. Scherer, COVAREP—A collaborative voice analysis repository for speech technologies, in: 2014 Ieee International Conference on Acoustics, Speech and Signal Processing, Icassp, IEEE, 2014, pp. 960–964, http://dx.doi.org/10.1109/ICASSP.2014.6853739.

[27] Y. Wang, Y. Shen, Z. Liu, P.P. Liang, A. Zadeh, L.-P. Morency, Words can shift: Dynamically adjusting word representations using nonverbal behaviors, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, (no. 1) 2019, pp. 7216–7223, http://dx.doi.org/10.1609/aaai.v33i01.33017216.

[28] M. Chen, X. Li, Swafn: Sentimental words aware fusion network for multimodal sentiment analysis, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 1067–1077, http://dx.doi.org/10.18653/v1/2020.coling-main.93.

[29] S. Baccianella, A. Esuli, F. Sebastiani, et al., Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining, in: Lrec, Vol. 10, No. 2010, 2010, pp. 2200–2204.

[30] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/N19-1423.

[31] A. Zadeh, R. Zellers, E. Pincus, L.-P. Morency, Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, 2016, http://dx.doi.org/10.48550/arXiv.1606.06259, arXiv preprint arXiv:1606.06259.

[32] A.B. Zadeh, P.P. Liang, S. Poria, E. Cambria, L.-P. Morency, Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2236–2246, http://dx.doi.org/10.18653/v1/P18-1208.

[33] K.D. Joshi, P.S. Nalwade, Modified k-means for better initial cluster centres, Int. J. Comput. Sci. Mob. Comput. 2 (7) (2013) 219–223.

[34] W. Rahman, M.K. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, E. Hoque, Integrating multimodal information in large pretrained transformers, in: Proceedings of the Conference. Association for Computational Linguistics. Meeting, vol. 2020, NIH Public Access, 2020, p. 2359, http://dx.doi.org/10.18653/v1/2020.acl-main.214.

[35] G. Paraskevopoulos, E. Georgiou, A. Potamianos, Mmlatch: Bottom-up top-down fusion for multimodal sentiment analysis, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2022, pp. 4573–4577, http://dx.doi.org/10.1109/ICASSP43922.2022.9746418.

[36] A. Joshi, A. Bhat, A. Jain, A. Singh, A. Modi, COGMEN: COntextualized GNN based multimodal emotion recognition, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2022, pp. 4148–4164, http://dx.doi.org/10.18653/v1/2022.naacl-main.306.

[37] S. Han, J. Hessel, N. Dziri, Y. Choi, Y. Yu, CHAMPAGNE: Learning real-world conversation from large-scale web videos, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2023, pp. 15498–15509, http://dx.doi.org/10.1109/iccv51070.2023.01421.

[38] T. Sun, J. Ni, W. Wang, L. Jing, Y. Wei, L. Nie, General debiasing for multimodal sentiment analysis, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 5861–5869, http://dx.doi.org/10.1145/3581783.3612051.

Zijun Liu was born in Nanchang, Jiangxi, China in 1999. In 2021, she entered the School of Software Engineering of Yunnan University (YNU) in China to study for a master's degree in software engineering. Her research interests include natural language processing, deep learning and emotion classification.

Li Cai was born in Kunming, China, in 1975. She received the M.S. degree in computer application from Yunnan University, China, in 2007. Then, she received the Ph.D. degree with the School of Computer Science, Fudan University, China, in 2020. From 1997 to 2002, she was a Research Assistant with Network Center. Since 2010, she has been an Associate Professor with the School of Software, Yunnan University. Her research interests include intelligent transportation, machine learning, visualization, and data quality.

Wenjie Yang was born in 1978 in Kunming, Yunnan, China. She received her M.S. in Human-care Science in 2009 from the University of Tsukuba(Japan), and Ph.D. degree in the same field from the same university in 2022. she was a lecturer in the Mental Health Counseling Center of Yunnan University from 2009 to 2022. Since October 2022, she has been an Associate Professor at Yunnan University. Her research interests include mental health education, counseling, and psychotherapy.

Junhui Liu received the M.S. degree in Software Engineering and the Ph.D. degree in Systems Analysis and Integration from Yunnan University (YNU), Kunming, China, in 2005 and 2009, respectively. He is currently a Lecturer with the School of Software, Yunnan University, Kunming, China. His current research interests include domain-specific modeling, deep reinforcement learning, time series forecasting, and machine learning.