International Journal of Science and Engineering Applications
Volume 14-Issue 10, 126 – 135, 2025, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1410.1019

Adaptive Learning Multimodal Sentiment Analysis based on Transformer

Shu Wei-Liang
Yangtze University
Jingzhou, Hubei, China

Abstract: Multimodal sentiment analysis aims to integrate data such as text, audio and video to get accurate sentiment predictions. Existing methods mainly focus on fusing information from multimodal data. However, modality-specific heterogeneity features interact , and irrelevant and conflicting information across modalities may hinder further performance improvement, for which a Transformer-based adaptive learning multimodal sentiment analysis model is proposed. Firstly, three multilayer perceptrons are utilized to extract multimodal features from multimodal data, then the Transformer separation encoder is introduced to de-entangle the inter-modal sentiment consistency features from the sentiment heterogeneity features, the textual modality among the heterogeneity features is dominated to do the attentional reinforcement to the audio and the video, and finally, the bidirectional sentiment query learning method is designed for the fusion of the sentiment features, Thus it provides a more comprehensive and complementary multimodal emotional representation, enhancing the capability of multimodal emotion recognition. This method was applied to three popular multimodal datasets, CH-SIMS, CMU-MOSI, and CMU-MOSEI. The results showed that the F1 value reached 82.31% on CH-SIMS, which was 0.6% higher than the second-best model; on CMU-MOSI, the F1 value reached 86.55%, which was 0.5% higher than the second-best model; and on CMU-MOSEI, the F1 value reached 86.64%, which was 0.4% higher than the second-best model. A large number of experiments have proved that the model proposed in this paper has certain competitiveness compared with other models.

Keywords: multimodal; sentiment analysis; multimodal sentiment disentanglement; modal interaction; feature fusion; adaptive learning

1. INTRODUCTION

Sentiment analysis, also known as opinion mining, is an artificial intelligence (AI) technique that analyzes, processes, summarizes, and infers emotions from subjective textual data [1-2].

Early sentiment analysis primarily focused on text-based data, but relying solely on text cannot fully capture the complexity of human emotional expression. For instance, while a model might classify the word "excellent" as generally positive, its meaning can shift to negative sentiment when accompanied by exaggerated or sarcastic facial expressions or tone. To address this limitation, multimodal sentiment analysis has emerged, integrating multiple data sources (e.g., text, voice, and facial expressions) to achieve more accurate and nuanced emotion recognition [3].

Multimodal sentiment analysis (MSA) is a crucial yet challenging task in natural language processing (NLP), which aims to infer an opinion holder's attitude or perspective by integrating multiple modalities, such as text, video, and audio data [4].

(a) Enjoying life: Crispy spring rolls (Positive)

(b) We're vacationing at the beach (Positive)

Fig.1 Example of multimodal data

As illustrated in Figure 1(a), the textual expressions (e.g., "enjoying" and "crispy") convey positive sentiment, while the accompanying image further enhances the emotional interpretation by providing vivid visual context. In contrast, Figure 1(b) demonstrates a case where the text alone lacks clear emotional cues, yet the serene landscape imagery supplements the textual data, enabling a more accurate prediction of positive sentiment.

Sentiment analysis has emerged as a key research focus in artificial intelligence and affective computing [8-10]. Compared to traditional unimodal sentiment analysis, which relies solely on a single data source [11], multimodal sentiment analysis (MSA)—leveraging diverse modalities such as video, audio, and text—demonstrates significant advantages in enhancing emotional understanding and better aligns with real-world affective interactions and applications [12].

The core challenge of MSA lies in multimodal fusion. Sun et al. [8] introduced deep canonical correlation analysis to capture cross-modal correlated features. Liang et al. [13] developed a model that addresses distribution discrepancies in modality-invariant spaces, thereby mitigating inter-modal heterogeneity. Zadeh et al. [14] proposed a tensor fusion network, dynamically modeling interactions by embedding multimodal features into a tensor space, successfully overcoming fusion challenges in MSA. While these advances excel at modeling shared emotional information through fusion techniques, they often overlook modality-specific characteristics, limiting their ability to effectively model nuanced cross-modal emotional relationships and ultimately hindering performance.

Transformer-based architectures, renowned for their relational learning and sequence modeling capabilities, have been widely adopted in computer vision, NLP, and MSA [15-16]. For instance, Delbrouck et al. [17] introduced a novel Transformer-based encoder with modular attention mechanisms to achieve superior multimodal fusion. Wang et al. [18] proposed a lightweight attention aggregation module combined with a cross-modal Transformer, employing gated recurrent units and enhanced feature extraction for low-resource modalities. Han et al. [19] designed an end-to-end network to fuse modalities by modeling their relationships with emotions. Zhuang et al. [20] improved Transformer-based video sequence modeling to extract higher-level feature abstractions. Despite their effectiveness in capturing cross-modal emotional dependencies, standard Transformers fail to fully account for the subtle heterogeneities in modality-emotion interactions, restricting further performance gains.

To address this, adversarial learning has been explored to disentangle shared emotional representations from modality-specific features. Yuhao et al. [21] employed adversarial training to decouple domain-invariant sentiment representations, facilitating cross-domain emotion transfer. He et al. [22] reduced inter-modal heterogeneity via adversarial encoders to derive modality-invariant features. Jie et al. [23] proposed a multimodal interaction model with adversarial training to learn relationships between text, images, and aspects. Nevertheless, the intricate and nuanced interplay of heterogeneous information across modalities remains a critical challenge, impeding comprehensive multimodal representation learning. For example, prosodic emotional cues in audio lack direct counterparts in text or video.

Current approaches face two key limitations:

Standard Transformers struggle to model subtle cross-modal emotional interactions, often overlooking discriminative features.

Adversarial methods require meticulously designed modules and extensive training data, risking model overcomplexity and reduced robustness.

Future work should focus on lightweight, adaptive fusion mechanisms that balance shared and modality-unique information while maintaining computational efficiency.

To address the aforementioned issues, this paper first employs three multilayer perceptrons (MLPs) to extract multimodal features from the input data. We then adopt a parameter-shared disentangled encoder to separately model both inter-modal emotional consistency and intra-modal emotional heterogeneity. Subsequently, we introduce an attention enhancement module that effectively guides the visual and auditory modalities to filter out emotion-irrelevant information through multi-scale linguistic features. Finally, to explore the emotional interactions between disentangled features, we design an interactive Transformer that integrates the disentangled features by performing bidirectional query learning across two cross-modal encoders, thereby obtaining more comprehensive multimodal sentiment representations. The main contributions of this paper are as follows:

(1) We propose a novel Transformer-based adaptive learning framework for multimodal sentiment analysis, which introduces two key innovations: a multimodal sentiment disentanglement mechanism and an interactive Transformer module. These components work synergistically to effectively extract and fuse emotional information across modalities, significantly enhancing the model's robustness.

(2) The linguistic features guide the learning of video and audio features to capture complementary heterogeneous representations that incorporate both cross-modal correlations and conflict-suppressing information, thereby reducing interference from redundant data.

(3) Extensive experiments conducted on benchmark datasets - including the Chinese dataset CH-SIMS [5] and English datasets CMU-MOSI [6] and CMU-MOSEI [7] - demonstrate that our method achieves superior performance in multimodal sentiment analysis, with significant improvements in classification accuracy compared to state-of-the-art approaches.

2. PAGE SIZE

All material on each page should fit within a rectangle of 18 x 23.5 cm (7" x 9.25"), centered on the page, beginning 2.54 cm (1") from the top of the page and ending with 2.54 cm (1") from the bottom. The right and left margins should be 1.9 cm (.75”). The text should be in two 8.45 cm (3.33") columns with a .83 cm (.33") gutter.

3. METHODS

The proposed Transformer-based Adaptive Learning for Multimodal Sentiment Analysis (TAL-MSA) framework, as illustrated in Figure 2, comprises four key components: (1) feature extraction, (2) sentiment disentanglement, (3) adaptive modality learning, (4) interactive Transformer.

First, we extract multimodal features (text, audio, video) from the input data. Next, we introduce a Transformer-based disentangled encoder to separate cross-modal sentiment-consistent features from modality-specific heterogeneous features. Building upon this, we focus on the textual modality within the heterogeneous features as the dominant cue to compute attention-weighted representations for the audio and visual features through a guided attention mechanism. Finally, we design a bidirectional affective query learning approach to effectively integrate these diverse emotional features, generating a comprehensive multimodal fusion vector for sentiment prediction.

3.1 3.1 Multimodal Feature Extraction

In multimodal sentiment analysis, pre-computed features are conventionally employed as input data across different modalities [24]. To align the feature vectors from three modalities, we utilize three parallel Multilayer Perceptrons (MLPs) [25] to normalize the sequence dimensions of each feature vector to a unified dimension d .

In this paper, let $U_t \in R^{T_o \times d_t}$,$U_a \in R^{T_o \times d_a}$ and $U_v \in R^{T_o \times d_v}$ denote the pre-computed feature vectors for text, audio, and video modalities respectively, where T_o represent the sequence lengths of text, audio, and video features, and d_t, d_a, d_v indicate the dimensional sizes of text, audio, and video feature vectors. The concatenation operation is formally defined in Equation (1):

$$U_0 = con\left(MLP\left(U_a\right), MLP\left(U_v\right), MLP\left(U_t\right)\right) \quad (1)$$

thereby obtaining a multimodal feature vector U_0 with $3T_o \times d$ dimensions.

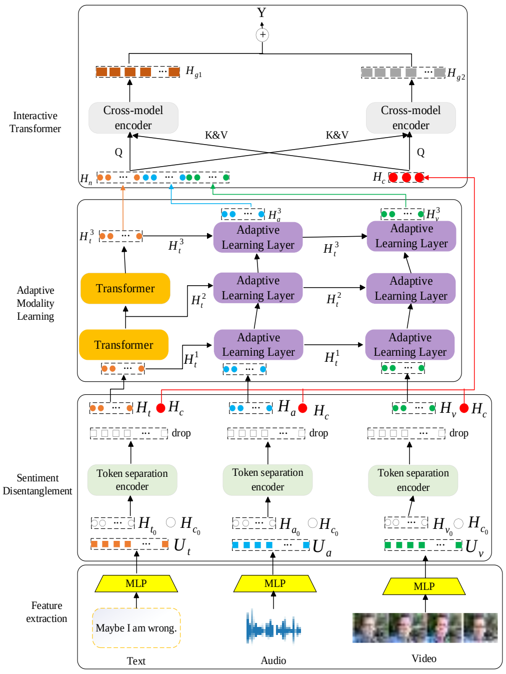

Fig.2 Adaptive learning multimodal sentiment analysis model based on Transformer

3.2 3.2 Multimodal Sentiment Disentanglement

The multimodal sentiment disentanglement module employs a parameter-shared Transformer encoder as the disentangling encoder to separate multimodal consistent features from heterogeneous features. The disentanglement learning process is illustrated in Figure 2.

The feature vectors H_{c_0}, H_{c_0}, H_{a_0}, and H_{v_0} are randomly initialized. Let H denote the feature vector, where the subscript c represents multimodal sentiment-consistent features, and subscripts t, a, and v denote modality-specific heterogeneous features for text, audio, and video modalities respectively. The subscript 0 indicates randomly initialized feature vectors. The dimension of H_{c_0} is $6\times d$, while H_{a_0}, H_{t_0}, and H_{v_0} all have dimension $2\times d$.

To replace the single input tensor in the Transformer, we concatenate H_{c_0}, H_{t_0}, H_{a_0}, H_{v_0}, and U_0 row-wise to form the composite tensor $con\left(H_{c_0},H_{a_0},H_{v_0},H_{t_0},U_0\right)$, where $con(\ )$ represents the concatenation operation. The concatenated tensor is then fed into the disentangling encoder $\mathrm{T}(\ )$ to extract four disentangled sentiment features: sentiment-consistent feature H_c and sentiment-heterogeneous features H_t, H_a, and H_v. This process is formally described in Equation (2):

$$H_c,H_a,H_v,H_t = \mathrm{T}\left(q/k/v = con\left(H_{c_0},H_{a_0},H_{v_0},H_{t_0},U_0\right)\right) \qquad (2)$$

Let q, k, and v denote the input tensors in the Transformer encoder, where $\left(3T_o+12\right)\times d$ specifies the tensor dimensionality.

Following the canonical Transformer encoder architecture [26], $\mathrm{T}(\ )$ comprises multi-head attention mechanisms, normalization layers, and multilayer perceptrons (MLPs), as depicted in Figure 3. Additionally, we adopt learnable positional encodings to process temporal sequence information, which are additively combined with the sequence content to form the integrated input representation for the Transformer. This approach enables the Transformer to process both content and positional information simultaneously yet distinctly, preventing their mutual interference and thereby significantly enhancing the model's capacity for sequential structure comprehension.

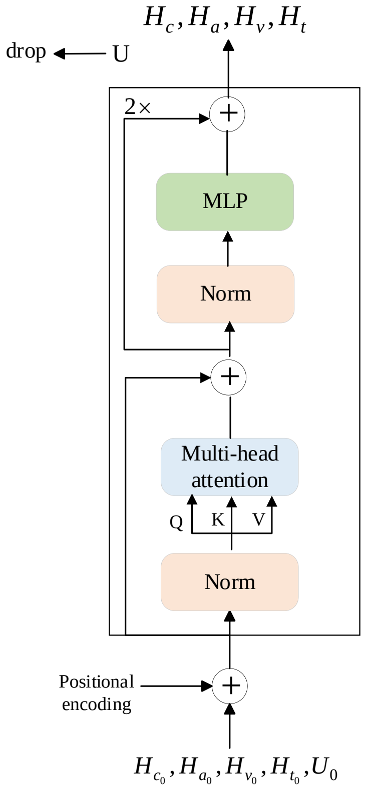

Fig.3 The structure of Transformer

3.3 Adaptive Modality Learning

Following the disentanglement process, we further employ an adaptive modality learning module to learn heterogeneous feature representations that capture cross-modal correlations, suppress inter-modal conflicts, and incorporate linguistically complementary information. As shown in Figure 2, this module consists of two Transformer layers and three adaptive learning layers for feature refinement. The adaptive learning layers dynamically adjust feature contributions through modality-attentive weighting, conflict-aware suppression, and correlation-enhanced fusion mechanisms.

Text Feature Construction

The text feature construction process is formally defined by Equation (3):

$$H_t^{i+1} = E_t^i\left(H_t^i, \theta_{E_t^i}\right) \in R^{T\times d} \tag{3}$$

Where $i \in \{1,2\}$, E_t^i and $\theta_{E_t^i}$ denote the i -th Transformer layer and its corresponding parameters for learning textual features, employing an 8-head attention mechanism to model information across each channel. This yields H_t^2 and H_t^3, representing medium- and high-scale text features respectively, while the initial feature H_t^1 serves as the low-scale text feature. All three features share the same dimensionality $T \times d$.

3.3 Adaptive Learning Layer

The adaptive learning layer is illustrated in Figure 4. To address the complex and nuanced heterogeneous information interactions across modalities, we utilize textual features as guidance to enhance effective fusion and reinforcement of video and audio features through attention mechanisms. Specifically, we first employ multi-scale text feature H_t^i as the query, while using H_v and H_a as the key and value respectively. This process computes the similarity matrix α between text and video features, as formalized in Equation (4):

$$\alpha = \text{softmax}\left(\frac{Q_t K_v^T}{\sqrt{d_k}}\right) = \text{softmax}\left(\frac{H_t^i W_{Q_t} W_{K_v}^T H_v^{1T}}{\sqrt{d_k}}\right) \in R^{T\times T} \tag{4}$$

In the attention mechanism, the softmax function performs weight normalization. Here, $W_{Q_t} \in R^{d\times d_k}$ and $W_{K_v} \in R^{d\times d_k}$ are two learnable parameter matrices, while d_k denotes the dimensionality of each attention head, which is specifically set to 16.

Similarly, the cross-modal similarity matrix β for text-audio feature interactions is formally defined in Equation (5):

$$\beta = \text{softmax}\left(\frac{Q_t K_a^T}{\sqrt{d_k}}\right) = \text{softmax}\left(\frac{H_t^i W_{Q_t} W_{K_a}^T H_a^{1T}}{\sqrt{d_k}}\right) \in R^{T\times T} \tag{5}$$

$W_{K_a} \in R^{d\times d_k}$ is a learnable parameter matrix.

Subsequently, the feature representations of these two modalities are updated using the weighted video and audio features, as formalized in Equations (6) and (7):

$$H_v^i = \alpha V_v = \alpha H_v^{i-1} W_{V_v} \tag{6}$$

$$H_a^i = \beta V_a = \beta H_a^{i-1} W_{V_a} \tag{7}$$

Here, $i \in (1,2,3)$. Within the adaptive learning layer, $H_v^i \in R^{T\times d}$ and $H_a^i \in R^{T\times d}$ represent the output features of the video and audio modalities from the i -th layer respectively, while $W_{V_v} \in R^{d\times d_k}$ and $W_{V_a} \in R^{d\times d_k}$ denote learnable parameter matrices.

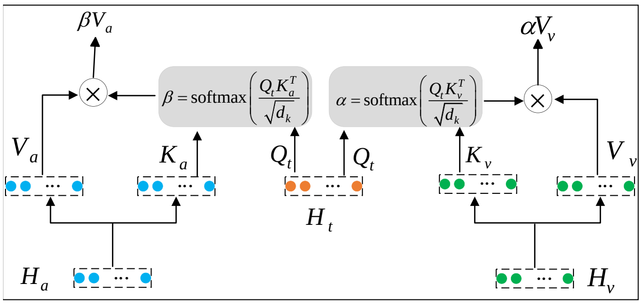
βV_a
αV_v
$\beta = \text{softmax}\left(\frac{Q_t K_a^T}{\sqrt{d_k}}\right)$
$\alpha = \text{softmax}\left(\frac{Q_t K_v^T}{\sqrt{d_k}}\right)$
V_a
K_a
Q_t
Q_t
K_v
V_v
H_t
H_a
H_v

Fig.4 Adaptive learning layer

3.4 3.4 Interactive Transformer

Current Transformer-based methods primarily focus on query representations, which often leads to over-emphasis on the emotional characteristics of the query modality while neglecting the complex and nuanced cross-modal affective interactions. This limitation frequently results in incomplete fusion of emotional features. To address this constraint, we design an Interactive Transformer module that achieves superior sentiment analysis performance through bidirectional query learning across two cross-modal encoders.

The modality-specific features H_t^3, H_a^3 and H_v^3 obtained from the adaptive learning module are concatenated to construct the composite heterogeneous feature H_n, as formally defined in Equation (8):

$$H_n = \{con\left(H_t^3, H_a^3, H_v^3\right)\} \tag{8}$$

We employ two parallel cross-modal encoders, denoted as $\mathrm{T}_{c\to n}(\ )$ and $\mathrm{T}_{n\to c}(\ )$, to implement a bidirectional query learning strategy for deeply integrating features H_n and H_c.

To investigate the contribution of heterogeneous features to affective interactions, we employ heterogeneous feature H_n as queries, consistent feature H_c as keys and values, and $\mathrm{T}_{c\to n}(\ )$ for heterogeneity-to-consistency sentiment fusion. During the transformation, an 8-head attention mechanism is adopted to model relationships and perform feature integration. The learning process of $\mathrm{T}_{c\to n}(\ )$ is formally expressed in Equation (9):

$$H_{g1} = \mathrm{T}_{c\to n}\left(q = H_n, k/v = H_c\right) \tag{9}$$

H_{g1} represents the fused feature mapping from heterogeneous to consistent representations.

To explore the contribution of consistent features to affective interactions, we employ consistent feature H_c as queries, heterogeneous feature H_n as keys and values, and $\mathrm{T}_{n\to c}(\ )$ for consistency-to-heterogeneity sentiment fusion. The learning process of $\mathrm{T}_{n\to c}(\ )$ is formally defined in Equation (10):

$$H_{g2} = \mathrm{T}_{n\to c}\left(q = H_c, k/v = H_n\right) \qquad (10)$$

H_{g2} is the consistency-to-heterogeneity fusion feature.

Through bidirectional query learning in two cross-modal encoders, we comprehensively capture subtle affective interactions from complementary perspectives, yielding two fused emotional representations H_{g1} and H_{g2}. Finally, the summation of H_{g1} and H_{g2} produces the ultimate multimodal sentiment prediction.

3.5 Loss Function

In the proposed method, the mathematical formulation of the adopted loss function is specified in Equation (11):

$$\ell = \frac{1}{N_b}\sum_{n=0}^{N_b}\left\|y^n - \hat{y}^n\right\|_2^2 \qquad (11)$$

N_b denotes the total number of samples in the training set, y^n represents the ground-truth sentiment label of the n-th sample, and $\hat{y}^n$ indicates our model's predicted value for that sample.

4. RESULT

4.1 Dataset

This paper conducts experiments on three public benchmark datasets: CH-SIMS, CMU-MOSI, and CMU-MOSEI.

CH-SIMS is a Chinese multimodal sentiment analysis dataset with fine-grained modality annotations, covering 2,281 carefully selected video clips, each equipped with both multimodal and independent unimodal annotations. The labels range from -1 to 1, divided into three sentiment categories: negative, neutral and positive.

CMU-MOSI is a widely used multimodal sentiment analysis dataset consisting of 89 videos from 89 speakers, containing 2,199 video clips in total. Each clip is manually annotated with sentiment scores ranging from -3 (strongly negative) to 3 (strongly positive), covering seven levels.

CMU-MOSEI is a comprehensive multimodal sentiment analysis dataset developed by researchers at Carnegie Mellon University, containing 23,453 annotated video segments from 1,000 different speakers and 250 topics. The dataset provides both sentiment and emotion labels, with emotions categorized into six classes: anger, happiness, sadness, surprise, fear and disgust. The emotion intensity is annotated on a continuous scale from -3 (strongly negative) to 3 (strongly positive).

4.2 Implementation Details And Evaluation Criteria

The experiments described in this paper are implemented using the PyTorch 1.11.0 deep learning framework on an NVIDIA GeForce RTX 4060 Ti GPU. To evaluate model performance, we employ the following metrics: for the CH-SIMS dataset, binary classification accuracy (acc_2), weighted F1 score (F1), three-class accuracy (acc_3), five-class accuracy (acc_5), mean absolute error (MAE), and Pearson correlation coefficient (Corr); for the CMU-MOSI and CMU-MOSEI datasets, binary classification accuracy (acc_2), weighted F1 score (F1), seven-class accuracy (acc_7), mean absolute error (MAE), and Pearson correlation coefficient (Corr). In our implementation, each layer of the Transformer encoder is configured with an 8-head attention mechanism, multilayer perceptrons, and 2 normalization layers. All datasets use a batch size of 64 and the Adam optimizer with a learning rate of 0.0001. The dimensionality of sentiment-consistent features is 6×256, while that of sentiment-heterogeneous features is 2×256.

4.3 Baselines

The experimental comparisons in this study selected the following baseline models:

TFN [14]: The Tensor Fusion Network (TFN) comprises modality embedding sub-networks, a tensor fusion layer, and a sentiment inference sub-network, capturing inter-modal dynamics through tensor fusion.

MulT [27]: This model employs directional pairwise cross-modal attention to focus on interactions between multimodal sequences at different timesteps, potentially adjusting information flow from one modality to another.

ICCN [8]: The Interactive Canonical Correlation Network (ICCN) utilizes deep canonical correlation analysis to learn hidden relationships among text, audio, and video modalities to enhance multimodal language analysis.

MAG-BERT [28]: This approach captures visual and auditory information in vector form without altering the original model architecture, effectively fine-tuning BERT and XLNet for multimodal language data.

MISA [29]: By learning both modality-invariant and modality-specific representations, this model facilitates effective fusion through two subspace learning components: one for capturing shared features and another for modality-specific characteristics.

Self-MM [30]: Integrating a pretrained BERT model for text feature extraction, this method combines audio and visual information while employing a self-supervised learning strategy to learn modality-specific representations.

MMIM [31]: This model improves multimodal fusion through hierarchical mutual information maximization, preserving task-relevant information from input to fusion output.

ALMT [32]: Incorporating an adaptive hypermodality learning module, this approach learns representations that suppress irrelevant and conflicting information from visual and audio features under linguistic guidance at different scales.

TMBL [33]: The proposed Two- and Three-Modal Binding Learning model addresses how to effectively extract both modality-invariant and modality-specific features before fusion through combined bimodal and trimodal binding mechanisms.

4.4 Performance Comparison

The comparative results between the TAL-MSA model and the aforementioned baseline models on the CH-SIMS, CMU-MOSI, and CMU-MOSEI datasets are presented in Table X, with Table 1 showing the results on CH-SIMS, Table 2 on CMU-MOSI, and Table 3 on CMU-MOSEI. The "-" symbol indicates unavailable data.

The Transformer-based adaptive learning approach demonstrates significant performance improvements across all three datasets compared to other models. On CH-SIMS, it achieves a 0.5% higher binary classification accuracy (acc_2), 0.3% higher three-class accuracy (acc_3), 0.4% higher five-class accuracy (acc_5), 1.2% higher F1-score, 0.7% lower mean absolute error (MAE), and 0.6% higher Pearson correlation coefficient (Corr) than the second-best model. For CMU-MOSI, the improvements include 0.3% higher acc_2, 0.9% higher seven-class accuracy (acc_7), 0.5% higher F1-score, and 0.3% lower MAE. On CMU-MOSEI, the model outperforms the second-best approach by 0.3% in acc_2, 1% in acc_7, 0.4% in F1-score, and 0.2% in MAE.

Our model surpasses all baselines in most metrics, demonstrating substantial advantages. Compared to TFN and MulT, TAL-MSA achieves more accurate cross-modal representations through its sensitive capture and adaptive processing of inter-modal variations. Relative to MISA, TAL-MSA simplifies model complexity via disentanglement while obtaining more comprehensive representations. When compared with ICCN, ALMT, MAG-BERT, and Self-MM, TAL-MSA fully disentangles and fuses both inter-channel sentiment-consistent and intra-channel sentiment-heterogeneous features, using multi-scale linguistic features to guide the removal of irrelevant information in visual and auditory modalities, thereby enabling deeper mutual learning. Against MMIM and TMBL, TAL-MSA's Interactive Transformer further enhances cross-modal feature fusion, leading to superior sentiment analysis accuracy.

Table1 Experimental results on the CH-SIMS dataset

Model	acc_2 (↑)	acc_3 (↑)	acc_5 (↑)	F1 (↑)	MAE (↓)	Corr (↑)
TFN	78.38	65.12	39.70	78.62	0.432	0.591
MulT	78.84	67.13	38.24	79.66	0.453	0.564
MAG-BERT	74.44	-	-	71.75	0.492	0.399
MISA	80.10	64.99	41.79	76.59	0.447	0.563
Self-MM	80.04	65.47	41.53	80.44	0.425	0.595
MMIM	78.30	65.52	42.13	78.26	0.423	0.597
ALMT	80.39	66.93	44.73	81.10	0.417	0.609
TAL-MSA	**80.86**	**67.37**	**45.07**	**82.31**	**0.410**	**0.615**

Table2 Experimental results on the CMU-MOSI dataset

Model	acc_2 (↑)	acc_7 (↑)	F1 (↑)	MAE (↓)	Corr (↑)
TFN	80.81	34.97	80.74	0.901	0.698
MulT	83.04	40.05	82.88	0.871	0.698
ICCN	83.02	39.01	83.09	0.860	0.710
MAG-BERT	84.43	43.62	84.61	0.727	0.781
MISA	83.46	42.34	83.69	0.783	0.761
Self-MM	86.04	45.82	86.03	0.713	0.798
MMIM	86.06	46.65	85.98	0.700	0.801
ALMT	86.13	48.52	86.03	0.701	0.805
TMBL	84.36	36.30	84.34	0.762	**0.867**

TAL-MSA	**86.45**	**49.42**	**86.55**	**0.697**	0.803

Table3 Experimental results on the CMU-MOSEI dataset

Model	acc_2（↑）	acc_7（↑）	F1（↑）	MAE（↓）	Corr（↑）
TFN	82.55	50.24	82.12	0.593	0.700
MulT	82.59	51.87	82.34	0.580	0.703
ICCN	84.20	51.62	84.22	0.565	0.713
MAG-BERT	84.82	52.67	84.71	0.543	0.755
MISA	85.55	52.23	85.33	0.555	0.756
Self-MM	85.27	53.54	85.36	0.530	0.765
MMIM	85.00	53.23	85.17	0.536	0.764
ALMT	86.09	53.58	86.26	0.529	0.770
TMBL	85.84	52.45	86.12	0.523	**0.780**
TAL-MSA	**86.41**	**54.51**	**86.64**	**0.521**	0.772

4.5 Ablation Study

To comprehensively evaluate the effectiveness of the TAL-MSA model, this paper conducts ablation experiments on the CMU-MOSEI dataset. As shown in Table 4, the experimental results include tests using bimodal inputs only, as well as configurations with the disentanglement module removed, consistent features eliminated, heterogeneous features discarded, the adaptive learning module ablated, and the Interactive Transformer module replaced with a single Transformer. The results demonstrate the individual contributions of each component to the overall model performance.

Table4 Results of ablation experiments

Method	acc_2 （↑）	acc_7 （↑）	F1 （↑）	MAE （↓）	Corr （↑）
TAL-MSA	86.41	54.51	86.64	0.521	0.772
A+V	73.32	31.96	76.50	0.981	0.684
T+A	76.27	34.57	78.06	0.939	0.695
T+V	77.32	34.87	77.39	0.947	0.698
w/o disentanglement	82.60	52.12	83.98	0.570	0.761
w/o consistent features	80.35	50.27	84.30	0.783	0.750
w/o heterogeneous features	81.63	51.52	84.20	0.686	0.754
w/o adaptive learning	83.05	50.26	83.93	0.602	0.742
only Transformer	84.16	52.78	85.15	0.591	0.766

The ablation studies in Table 4 show that removing any modality reduces model performance. The T+V and T+A combinations perform significantly better than A+V, proving text contains more emotional information. Replacing the disentanglement module with MLP decreases recognition performance, showing the module effectively learns different emotional patterns. Removing consistent or heterogeneous features also reduces accuracy.

For cross-modal attention, the adaptive learning module significantly improves performance, demonstrating text's effectiveness in guiding video and audio modalities. Removing the Interactive Transformer decreases performance, proving its importance in integrating multimodal features and enhancing cross-modal emotional interaction.

Overall, removing any module reduces performance metrics, verifying each component's importance and providing key guidance for improving multimodal sentiment analysis models.

5. Conclusion

To address the challenges of modality-specific heterogeneous interactions and cross-modal irrelevant/conflicting information in multimodal sentiment analysis, this paper proposes a Transformer-based adaptive learning fusion method. First, we present the overall framework of our model and introduce the benchmark datasets used for validation: CH-SIMS, CMU-MOSI, and CMU-MOSEI. Ablation studies ultimately confirm the core importance of each module for sentiment analysis tasks.

For future work, we will adopt advanced supervised learning methods to better utilize unlabeled data. We also plan to investigate the integration of multimodal sentiment analysis with large-scale foundation models to enhance generalization capabilities.

6. References

[1] GIANNI B,FLAVIUS F. A survey on aspect-based sentiment classification[J].ACM Computing Surveys, 2022,55(4):1-37.

[2] ERIC C. Affective computing and sentiment analysis[J]. IEEE Intelligent Systems,2016,31(2):102-107.

[3] .CHEN G W, ZHANG P Z, WANG T, et al.Review on multimodal sentiment recognize[J],2022,29(2):70-78.

[4] SOLEYMANI M, DAVID G, JOU B, et al. A survey of multimodal sentiment analysis[J].Image and vision computing,2017,65:3-14.

[5] YU W, XU H, MENG F, et al. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality[C]//Proceedings of the 58th annual meeting of the association for computational linguistics. ONLINE: ACL Press,2020:3718-3727.

[6] ZADEH A, AELLERS R, PINCUS E, et al. Multimodal sentiment intensity in videos:facial gestures and verbal messages[J].IEEE Intelligent Systems,2016,31(6):82-88.

[7] ZADEH A B,LIANG P P,PORIA S, et al. Multimodal language analysis in the wild:Cmu-mosei dataset and interpretable dynamic fusion graph[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: ACL Press,2018:2236-2246.

[8] SUN Z, SARMA P, Sethares W, et al. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis[J].Proceedings of the AAAI Conference on Artificial Intelligence,2020:8992-8999.

[9] AN J, WAN ZAINON, W M N. Integrating color cues to improve multimodal sentiment analysis in social media[J]. Engineering Applications of Artificial Intelligence,2023,126(A):10.

[10] SINGH N, KAPOOR R. Multi-modal expression detection(MED): A cutting-edge review of current trends, challenges and solutions[J]. Engineering Applications of Artificial Intelligence,2023,125(C):23.

[11] CHEN Y, JOO J. Understanding and mitigating annotation bias in facial expression recognition[J]. arXiv preprint arXiv:2108.08504,2021.

[12] CHEN Z H.Multimodal Sentiment Analysis Based on DeepFusion Learning[D].XI'AN:XIDIAN UNIVERSITY,2023.

[13] LIANG T, LIN G, FENG L, et al. Attention is not enough: Mitigating the distribution discrepancy in asynchronous multimodal sequence fusion[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision.Montreal,QC,Canada:IEEE press,2021:8128-8136.

[14] ZADEH A, CHEN M, PORIA S, et al. Tensor fusion network for multimodal sentiment analysis[J].arXiv preprint arXiv:1707.07250,2017.

[15] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929,2020.

[16] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[J]. arXiv preprint arXiv:2103.14030,2021.

[17] DELBROUCK J B, TITS N, BROUSMICHE M, et al. A transformer-based joint-encoding for emotion recognition and sentiment analysis[C]//Proceedings of the Second Grand-Challenge and Workshop on Multimodal Language.Seattle,USA: ACL press,2020:1-7.

[18] WANG L, WANG Y,WANG J.Cross-Modal Transformer Combination Model for Sentiment Analysis[J].Computer Engineering and Applications,2024,60(13):124-135.

[19] HAN W, CHEN H, GELBUKH A F, et al.Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis[C]//Proceedings of the International Conference on Multimodal Interaction. MonteEAI, QC,Canada:ACM press,2021:6-15.

[20] ZHUANG G H,YILIHAMU Y.Multimodal sentiment analysis based on STD-Transformer network[J].Laser Journal,2022,43(08):95-100.

[21] YUHAO Z, YING Z, WENYA G, et al.Learning Disentangled Representation for Multimodal Cross-Domain Sentiment Analysis[J]. IEEE transactions on neural networks and learning systems,2022,34(10):7956-7966.

[22] HE J, SU B, SHENG Z, et al.Adversarial invariant-specific representations fusion network for multimodal sentiment analysis[J].International Conference on Image, Signal Processing, and Pattern Recognition,2023,12707:930-942.

[23] JIE Z, JIABAO Z, XIANGJI J H, et al.MASAD:A large-scale dataset for multimodal aspect-based sentiment analysis[J].Neurocomputing,2021,455:47-58.

[24] MAO H, YUAN Z, XU H, et al.M-SENA:an integrated platform for multimodal sentiment analysis[J].arXiv preprint arXiv:2203.12441,2022.

[25] TAUD H, MAS J F. Multilayer perceptron (MLP)[J]. Geomatic approaches for modeling land change scenarios, 2018:451-455.

[26] VASWANI A.Attention is all you need[J].Advances in Neural Information Processing Systems,2017:6000-6010.

[27] TSAI Y H, BAI S, LIANG P P, et al. Multimodal transformer for unaligned multimodal language sequences[C]//Proceedings of the conference. Association for Computational Linguistics. Meeting. Florence, Italy: ACL Press, 2019:6558-6569.

[28] RAHMAN W, HASAN K, LEE S, et al.Integrating Multimodal Information in Large Pretrained Transformers.[C]//Meeting of the Association for Computational Linguistics. Online:Association for Computational Linguistics, 2020:2359-2369.

[29] HAZARIKA D, ZIMMERMANN R, PORIA S. MISA:Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis[C]//Proceedings of the 28th ACM international conference on multimedia.ONLINE:ACM Press,2020:1122-1131.

[30] YU W, XU H, YUAN Z, et al. Learningmodality-specific representations withself-supervised multi-task learning formultimodal sentimentanalysis[C]//Proceedings of the AAAIconference on artificial intelligence. ONLINE:AAAI Press, 2021: 10790-10797.

[31] HAN W, CHEN H, PORIA S. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis[J]. arXiv preprint arXiv:2109.00412, 2021.

[32] HAOYU Z, YU W, GUANGHAO Y, et al. Learning Language-guided Adaptive Hyper-modality Representation for Multimodal Sentiment Analysis[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing.Singapore: ACL Press,2023:756-767.

[33] HUANG J, ZHOU J, TANG Z, et al. TMBL:Transformer-based multimodal binding learning model for multimodal sentiment analysis[J].Knowledge-Based Systems, 2024,285: 111346