



Multimodal sentiment analysis based on multiple attention

Hongbin Wang, Chun Ren, Zhengtao Yu*

Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, Yunnan, PR China

Key Laboratory of Artificial Intelligence in Yunnan Province, Kunming University of Science and Technology, Kunming 650500, Yunnan, PR China

ARTICLE INFO

Dataset link: <https://twitter.com>, <http://tumblr.com>

Keywords:

Multimodal sentiment analysis

Multimodal interaction

Adaptive

Attention mechanism

ABSTRACT

The development of the Internet makes various types of data widely appear on various social platforms, multimodal data provides a new perspective for sentiment analysis. Although the data types are different, there are information expressing the same sentiment. The existing researches on extracting those information are static, and this means that there is a problem of extracting common information in a fixed amount. Therefore, to address this problem, we proposes a method named multimodal sentiment analysis based on multiple attention(MAMSA). Firstly, this method utilized the adaptive attention interaction module to dynamically determine the amount of information contributed by text and image features in multimodal fusion, and multimodal common representations are extracted through cross modal attention to improve the performance of each modal feature representation. Secondly, using sentiment information as a guide to extract text and image features related to sentiment. Finally, using hierarchical manner to fully learning the internal correlations between sentiment-text association representation, sentiment-image association representation, and multimodal common information to improve the performance of the model. We conducted extensive experiments using two public multimodal datasets, and the experimental results validated the availability of the proposed method.

1. Introduction

In recent years, social media's proliferation has established it as a vital tool for sharing and communication. Users increasingly leverage these platforms to express sentiments. Early sentiment analysis primarily involved text mining to understand people's opinions, sentiment tendencies, and attitudes toward factors such as services, products, individuals, organizations, and events. Nowadays, the content shared on social media now includes more than just text, it also expresses richer sentiment through visual and audio. Therefore, Multimodal Sentiment Analysis (MSA) has attracted widespread attention from researchers. Multimodal sentiment analysis can more accurately monitor users' sentimental responses to different topics, providing strong support for content recommendation and public opinion analysis on social media.

Multimodal sentiment analysis involves multiple data types, including speech, images, and text. Each data type has a unique way of expressing information, such as tones in speech, expressions in images, and vocabulary in texts. Multimodal sentiment analysis aims to harness correlations across modalities for more precise sentiment inference. Images and texts convey sentiment information, extracting those information poses a noteworthy challenge. Zhou et al. (2023a) used SenticNet to score each word and then added sentiment features to multimodal fusion features. Xiao et al. (2022) utilized a multi-head

self attention mechanism to integrate sentiment knowledge into inter modal feature representation. Zhu et al. (2023) proposed the sentiment Knowledge Enhanced Attention Fusion Network (SKEAFN), which enhances multimodal fusion by adding additional sentiment knowledge from external knowledge bases. These methods have proven that incorporating sentiment information into models has been proven effective in enhancing sentiment analysis accuracy, offering novel insights for our model design.

Analyzing multimodal sentiment necessitates attention to commonalities and complementarities across modalities. Fig. 1a shows a smiling face and "important" text conveying positive sentiment, information expressed diversely yet conveying the same sentiment is deemed common information between modalities. As shown in Fig. 1b, text sentiment is neutral, but the man's smile in the image suggests a positive sentiment of the image-text pair. This smile enhances the text, constituting complementary information. To tackle common information, several researchers have developed innovative methodologies in MSA. Hazarika et al. (2020) proposed a model called modality-invariant and-specific representations for multimodal sentiment analysis. It utilizes modal-invariant representations to capture common sentiment, minimizing cross-modal disparities. Similarly, Yang et al. (2022) proposed a method named FDMER for learning the public and private

* Corresponding author.

E-mail address: ztyu@hotmail.com (Z. Yu).



Fig. 1. Common information & complementary information.

feature representations of each modality, achieving modal consistency and difference constraints through customized loss functions. [Chen et al. \(2023\)](#) proposed a joint multimodal sentiment analysis method grounded in information correlation. Quantifies and models correlations between modalities to assess cross-modal feature matching at the sentiment level. While achieving notable results, these methods statically identify common information. In fact, the common information contained in each image-text pair is different, and extracting a fixed amount of common information is not applicable to all posts.

In summary, there is a problem with current research that extracting common information in a fixed amount cannot effectively adapt to all posts, resulting in inaccurate extraction of key information in some cases. So, we proposed a multimodal sentiment analysis method based on multiple attention. Instead of directly using extracted image-text features for sentiment analysis, it designs an adaptive attention interaction network. This network assigns distinct learnable parameters to features, adaptively determines text and image contributions in multimodal fusion, and extracts shared multimodal representations via cross-modal attention, enhancing each modal's feature representation. At the same time, in order to fully recognize the correlation between sentiments, text, and image, we also uses attention mechanisms to model sentiment-text association representations and sentiment-image association representations. Finally, hierarchical fusion is used to fully learn the internal correlation between sentiment-text association representation, sentiment-image association representation and multimodal common information in a hierarchical way, so as to improve the performance of the model.

The main contributions of this paper contain:

- We proposed a framework named multimodal sentiment analysis based on multiple attentions. The model employs an adaptive interaction module to dynamically determine the amount of common information and learn cross-modal commonalities. This module solves the problem of performance degradation caused by extracting a fixed amount of public information in existing research.
- We use sentiment attention to focus on information related to sentiment in text and images. Simultaneously design a hierarchical fusion module that fully interacts the three modalities using a layered approach to obtain internal correlations among them.
- Extensive experiments on two publicly available multimodal datasets have shown that our model can better focus on the common and complementary information between images and text. Compared to the previous baseline model, our model has shown significant improvement.

2. Related work

Acquiring the sentiment information of samples based on various perceptual channels such as text, image and speech, integrating them for sentiment analysis can undoubtedly reflect personal sentiment more comprehensively and greatly improve the accuracy. Firstly, the relationship between modalities is crucial for multimodal sentiment analysis. However, most of the existing multimodal sentiment analysis methods only splice the features of multiple modes together, unable to fully explore the interaction between them, resulting in unsatisfactory results. In addition, in the learning of multi-modal characteristics, the common information and complementary information between multimodal data are ignored, which leads to the deviation of sentiment analysis. Therefore, a large number of researchers have focused their attention on these two problems.

In order to achieve sufficient interaction and fusion between modalities, [Zhou et al. \(2023b\)](#) proposed a cross attention and mixed feature weighted network to achieve accurate sentiment recognition, fully utilizing the complementary information between image and contextual features. [Zhang et al. \(2021\)](#) used two memory networks to mine intra-modality information of images and texts, and then designed a discriminant matrix to supervise the fusion of inter-modality information. [Li et al. \(2022\)](#) proposed a multi-layer fusion model for aligning and fusing marker level features of text and images, and designed two comparative learning tasks to help the model learn sentiment related features in multimodal data. In order to fuse local features of image and text more effectively, they proposed MLF module that performs multimodal feature fusion from the fine-grained token-level. [Huang et al. \(2023\)](#) proposed a text-center fusion network with cross modal attention (TeFNA). [Liu et al. \(2023b\)](#) proposed a multimodal emotion recognition framework based on cascaded multi-channel hierarchical fusion (CMC-HF), which utilizes hierarchical fusion to effectively learn multi-modal information interaction. Similarly, [Yang et al. \(2021a\)](#) used stacked attention memory networks to make text features interact with image features, and constructed a multimodal feature fusion module using multi-layer perceptrons and stacked pool modules. Based on the latest advances in attention mechanisms, [Zhao et al. \(2023\)](#) proposed a shared private memory network called SPMN to decouple multimodal representations from private and shared perspectives. [Le et al. \(2023\)](#) proposed a fusion and representation learning method based on transformer, which takes the original video frames, audio signals, and text captions as inputs and transmits information from these multimodals through a unified transformer architecture to learn joint multimodal representations. [Zeng et al. \(2023\)](#) propose a Multimodal

Interactive and Fusion Graph Convolutional Network. Introducing image caption as an auxiliary to align with images to enhance semantic delivery. Then, use the generated sentences and images as nodes to construct a graph.

Different modalities express sentiments in different ways, but they all share the speaker's motivation and goals. Therefore, utilizing the commonality information between modalities will help analyze the sentiments of posts. For example, He et al. (2022) proposed a dynamic invariant specific representation fusion network that obtained joint domain separation representations of all modalities through an improved joint domain separation network, effectively utilizing fusion information. Liu et al. (2023a) proposed a knowledge distillation framework based on cross modal consistency modeling, which measures the semantic consistency of multimodal data by designing a hybrid course learning strategy. Xu et al. (2022) proposed the Multimodal Emotion Analysis Framework (CMJRT), which transfers joint representations from bimodal to unimodal through hierarchical interactions between modalities to obtain consistency and complementarity between modalities. The gated cross-modal attention mechanism proposed by Sun et al. (2023) performs modal interactions in an adaptive manner and filters out inconsistencies from multiple modes, and also uses parallel structures to learn more comprehensive affective information in a paired manner. Liu et al. (2023a) designed a hybrid curriculum learning strategy to measure semantic inconsistency in multimodal data, and gradually trained all image text pairs from easy to difficult, which can effectively handle the large amount of noise caused by inconsistent image and text on social media. Quan et al. (2022) proposed a new model framework, MICS, which adopts a suitable strategy for each modality and provides a better representation for fusion. Zhao et al. (2022) proposed a layered multimodal alignment and interactive network enhanced BERT. This model introduces a memory network to align different multimodal representations, and uses modal updating methods to address asynchrony issues. Lin et al. (2023) proposed a bidirectional style enhancement module to capture relevant semantic information between patterns. Xiao et al. (2023) proposed a cross modal fine-grained alignment and fusion network that converts images into text titles and graphic structures, dynamically aligning semantic and syntactic information from text titles and input text.

In summary, although these methods have achieved excellent results, existing research has ignored the importance of dynamically extracting common and complementary information between modalities, leading to bias in sentiment analysis. Therefore, we propose a model called multimodal sentiment analysis based on multiple attention, which designs an adaptive attention interaction network, assigns different learnable parameters to different features, and performs adaptive weighted feature attention interaction operations to extract multimodal common information. At the same time, in order to fully recognize the correlation between sentiments, text, and image, this model uses attention mechanisms to model sentiment-text association representations and sentiment-image association representations. Finally, multi-modal fusion is carried out in a hierarchical manner.

3. Proposed method

In this section, we will elaborate the details of our proposed model MAMSA, and the framework diagram of the model is shown in Fig. 2. The model is divided into five modules, which are feature encoding module, adaptive attention interaction module (AAI), sentiment association representation module (SA), hierarchical fusion module (HF), and sentiment prediction module.

3.1. Task definition

The purpose of MSA tasks is to obtain sentiments by using multi-modal signals of text (T) and image (P). Generally speaking, MSA can be seen as a classification task or regression task. we considers it as a classification task. Therefore, the model inputs text T_i and image P_i , and then outputs a sentiment label L_i .

3.2. Feature encoding

For text, a sentence can be composed of multiple words, $S_i = \{w_1, \dots, w_j, \dots, w_k\}$. We use Transformer stream of CLIP (Radford et al., 2021) to pre-train 400 million pairs of image-text pairs, and remove the last feature to encode the words in the sentence, Each word is encoded to 512-dimensional feature vector.

$$X_T^i = \text{TextTransformer}(S_i) \quad (1)$$

where $X_T^i \in \mathbb{R}^{k \times d}$, k represents the number of words in the sentence, $d = 512$ represents the dimension of the vector, and S_i represents the i_{th} sentence.

For image, as described in Vision Transformer (ViT) (Wu et al., 2021), an image can be represented by 16×16 words, in other words, an image is worth 16×16 patches. To represent n patches of an image P , we used ViT stream of CLIP, and remove the last feature choosing operation to encode each image patch. Finally, n patches of each image are encoded into 512-dimensional feature vector.

$$X_I^i = \text{VisionTransformer}(P_i) \quad (2)$$

where $X_I^i \in \mathbb{R}^{n \times d}$, n represents the number of patches, $d = 512$ represents the dimension of the vector, and P_i represents the i_{th} image.

3.3. Adaptive Attention Interaction (AAI)

After CLIP encoding, image and text features solely encompass their respective information. This paper posits that information from one modality can enhance the representation of the other, yet the extent of useful information for the opposing modality remains unknown. Hence, a network needs to be devised to dynamically determine this amount of information. At the same time, Simple concatenation between multiple modalities can bring more noise. To solve this problem, we design an adaptive attention interaction network, which assigns different learnable parameters to different features, and performing adaptive weighted feature attention interaction operations to extract common information.

Firstly, the text and image features encoded by Transformer are fused in the early, and the adaptive scores are dynamically obtained using the fused features. The scores represent the contribution degree of text and image during fusion, which helps to improve the performance of feature representation.

$$\widehat{X}_c^i = X_T^i \oplus X_I^i \quad (3)$$

$$\theta_i = \text{Sigmoid}(w_c^i \widehat{X}_c^i + b_c^i) \quad (4)$$

Multiplying the adaptive score with text features further determines the information contributed by text features in multimodal fusion, and similarly, by multiplying the $1 - \theta$ with the image features, the contribution of the image during fusion can be obtained. The specific process is as follows:

$$X_T^{i'} = \theta_i X_T^i \quad (5)$$

$$X_I^{i'} = (1 - \theta_i) X_I^i \quad (6)$$

where θ_i represents the contribution degree of text features X_T^i and image features X_I^i during fusion, Sigmoid is $\frac{1}{1+e^{-x}}$, which can map any real value to the (0,1) interval. w_c^i is a learnable weight, b_c^i is a learnable bias. $X_T^{i'}$ and $X_I^{i'}$ respectively represent text contribution information and image contribution information.

In order to strengthen the interaction between text contribution information and pictures, as well as the interaction between picture contribution information and text, a cross modal attention module is used to achieve this process. The implementation of the cross attention module CA is based on Scaled dot-product Cross-Attention (Vaswani et al., 2017), which can calculate the attention score between text

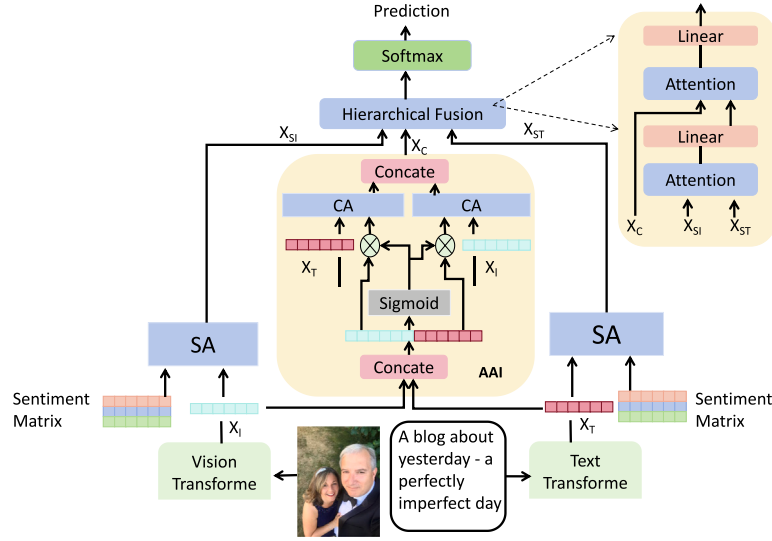


Fig. 2. The framework of MAMSA.

contribution information and images, as well as between image contribution information and text. Taking text features as an example, the query Q for cross modal attention is image contribution information X_I^i , the key K and value V are both text features X_T^i . The specific process is as follows.

$$s(X_I^i, X_T^i) = \frac{X_I^i \cdot X_T^i}{\sqrt{d}} \quad (7)$$

$$\alpha(X_I^i, X_T^i) = \text{softmax}(s(X_I^i, X_T^i)) \quad (8)$$

$$\tilde{X}_T^i = \alpha(X_I^i, X_T^i) X_T^i \quad (9)$$

$$\tilde{X}_I^i = \text{softmax}\left(\frac{X_T^i \cdot X_I^i}{\sqrt{d}}\right) X_I^i \quad (10)$$

where $s(X_I^i, X_T^i)$ is the dot product similarity function, $\alpha(X_I^i, X_T^i)$ is a weight score function that measures the importance of image contribution information X_I^i to text features X_T^i . \tilde{X}_T^i is a text representation related to image contribution information, denoted as a text common feature, and \tilde{X}_I^i is an image common feature.

Finally, the common information of the two modes is spliced to obtain the multimode common information:

$$X_C^i = \tilde{X}_T^i \oplus \tilde{X}_I^i \quad (11)$$

3.4. Sentiment Association representation (SA)

In order to fully recognize the correlation between sentiments, text, and image, we use attention mechanism to model the correlation representation between sentiments and text, and learn the text feature representation that is most relevant to sentiments. In addition, we also model the association representation between sentiments and image, focusing on the image regions that are most relevant to sentiments.

We take the sentiment embedding matrix (Yang et al., 2021b) as the query Q for Sentiment Text attention (STA), take the text feature as K and V of STA. That is, in the case of K = V, guided by sentiments, dynamically selects the word part that best expresses sentiment from semantic information, and generates a sentiment-text association representation.

For the establishment of the sentiment embedding matrix, we use the GLOVE model. For the MVSA-Single dataset, $L_i \in \{\text{positive, negative, neutral}\}$; For the TumEmo dataset, $L_i \in \{\text{angry, bored, clam, fear, happy, love, sad}\}$; Firstly, GLOVE is used to encode different words

to obtain word embedding vectors, which are filled to a specific length. Then, different embedding vectors are combined into an sentiment embedding matrix.

$$M_s = \text{GLOVE}(L), M_s \in \mathbb{R}^{3 \times d} \quad (12)$$

or

$$M_s = \text{GLOVE}(L), M_s \in \mathbb{R}^{7 \times d} \quad (13)$$

where among them, L represents the sentiment label, and $d = 300$ represents the dimensionality of the hidden layer representation.

$$X_{ST}^i = \text{softmax}\left(\frac{M_s \cdot X_T^i}{\sqrt{d}}\right) X_T^i \quad (14)$$

where M_s is the sentiment embedding matrix, $M_s \in \mathbb{R}^{3 \times d}$ when classifying sentiment polarity, $M_s \in \mathbb{R}^{7 \times d}$ when classifying sentiments. X_{ST}^i represents the sentiment-text association of the i th text-image pair.

Similarly, the sentiment embedding matrix is used as the query Q for sentiment Image Attention (SIA), take the image features as K and V of SIA. In the case of K = V, guided by sentiments, dynamic attention is paid to the regions in the image information that best express sentiment, generating an sentiment-image correlation representation.

$$X_{SI}^i = \text{softmax}\left(\frac{M_s \cdot X_I^i}{\sqrt{d}}\right) X_I^i \quad (15)$$

where X_{SI}^i represents the sentiment-image association of the i th text-image pair.

3.5. Hierarchical Fusion (HF)

After processing by AAI and SA, sentiment-text association representation, sentiment-image association representation, and multi-modal common information are obtained. In order to further integrate the three, we introduces a hierarchical fusion module. The design of hierarchical fusion is to unify and combine the three modalities ($X_{ST}^i, X_{SI}^i, X_C^i$) to obtain the internal correlation between the three. Early fusion and late fusion are unable to learn the internal correlations between heterogeneous modalities. Inspired by hybrid fusion strategies, hierarchical fusion fully utilizes the internal correlations between the three modalities in a hierarchical manner.

Hierarchical Fusion (HF) consists of two layers of Attention. In the first layer of Attention, Q is the sentiment-text association representation, while K and V are the sentiment-image association representation. This setting allows the model to focus on the image regions related to it when processing text, enabling the model to more accurately understand the image content and generate text output closely related to the image. After the first layer of Attention, we obtained an sentiment-multimodal association representation X_S^i . In the second layer of Attention, Q is the sentiment-multimodal association representation X_S^i , while K and V are the multimodal common information X_C^i . This setting can better focus on the sentiment information in the multimodal common information, and this process obtains the sentiment common multimodal representation to enhance our prediction accuracy. As shown in the upper right part of Fig. 2, the specific process is as follows:

$$X_S^{i'} = \text{Att}(Q = X_{ST}^i, K = X_{SI}^i, V = X_{SI}^i) \quad (16)$$

$$X_S^i = \text{Linear}(X_S^{i'}) \quad (17)$$

$$X_C^{i'} = \text{Att}(Q = X_S^i, K = X_C^i, V = X_C^i) \quad (18)$$

$$X^i = \text{Linear}(X_C^{i'}) \quad (19)$$

where $\text{Att}(\bullet)$ is the attention mechanism, $\text{Linear}(\bullet)$ is used to convert into specific

3.6. Sentiment classification

The above final representation mentioned is sent to the Softmax classifier to obtain the final sentiment label classification.

$$L_i = \text{Softmax}(W X_i + b) \quad (20)$$

where $M \in \mathbb{R}^{d \times L}$, $b \in \mathbb{R}^L$, L is the number of label categories. If it is an sentiment polarity classification, $L = 3$; If it is sentiment classification, $L = 7$. L_i is the predicted label of the i th text-image pair, and the loss function of the model adopts the standard cross entropy loss:

$$\text{Loss} = - \sum_{i=1}^N (L_{\text{true}}^i \log L_i + (1 - L_{\text{true}}^i) \log(1 - L_i)). \quad (21)$$

L_{true}^i is the true label of i th text-image pair, N representing the total number of training samples in the datasets.

4. Experiments

To validate the accuracy of the model in predicting sentiments, we conducted experiments on two publicly available multimodal sentiment analysis datasets, MVSA-Single (Niu et al., 2016) and TumEmo (Yang et al., 2021b). And the model is compared with some unimodal and multimodal approaches. Among them, three categories of sentiment polarity prediction tasks (positive, neutral, negative) were performed on MVSA-Single, and seven categories of sentiment prediction tasks (angle, bored, clam, fear, happy, love, sad) were performed on TumEmo. A series of ablation experiments were conducted to verify the effectiveness of each module, followed by experiments on the order of hierarchical fusion, and finally visualized the representation of sentiments on the image.

4.1. Datasets

MVSA-Single contains 5129 image-text pairs, which contains one sentiment annotation per data. TumEmo is a multimodal weakly supervised emotion dataset that contains 190 000 image-text pairs crawled from Tumblr, with each image-text pair labeled with a different emotion. In order to make a fair comparison, we processed the two datasets in accordance with Yang et al. (2021b), randomly dividing the train

Table 1
Statistics for different datasets.

Dataset	Train	Val	Test	Total
MVSA-Single	3608	451	452	4511
TumEmo	156,204	19,525	19,536	195,265

dataset, validation dataset, and test dataset in an 8:1:1 ratio, as shown in Table 1.

The distribution of each category in the dataset is shown in Fig. 3. As shown in the figure, the distribution of data among different categories in the dataset is unbalanced. Therefore, we perform random oversampling on the smallest category so as to reduce the influence of datasets imbalance on the experiments.

4.2. Parameter setting and evaluation indicators

we uses Adam (Kingma and Ba, 2014) to optimize the model implemented by PyTorch. The Batch Size of MVSA-Single and TumEmo is set to 32,64. The initial learning rate is set to $1e-4$, and the setting of the learning rate scheduler varies depending on the dataset. On the MVSA-Single dataset, step size = 1 and gamma = 0.8; on the TumEmo dataset, step size = 2, gamma = 0.8. All models were tested on NVIDIA 3090 GPUs.

The most commonly used sentiment analysis evaluation indicators are accuracy (ACC) and F1 value. ACC and F1 value are used to measure the effectiveness and performance of sentiment analysis systems. ACC refers to the proportion of samples correctly predicted by the model to the total sample size, while F1 value is used to comprehensively evaluate the performance of the model. The larger the value, the better the performance of the model.

$$ACC = \frac{TP}{N} \times 100\%. \quad (22)$$

$$F1 = 2 \times \frac{P \times R}{P + R} \times 100\%. \quad (23)$$

where TP is the number of correctly predicted samples, N is the total number of samples, P represents accuracy, which is the proportion of true positive classes predicted, R represents recall rate, which is the proportion of correctly predicted positive classes to true positive classes, $F1$ is the harmonic average of accuracy and recall.

4.3. Baseline model

In the paper, baseline models are classified into three categories, namely text model, image model and multimodal model. The results of CIGNN model are compared with baseline models.

Text model: CNN (Kim, 2014) and BiLSTM (Wang and Yang, 2020) are famous models for text classification. TGNN (Huang et al., 2019) is a text-level graph neural network for text classification. BiACNN (Lai et al., 2015) is a combination of CNN and BiLSTM with attention mechanism for text sentiment classification.

Image model: ResNet (He et al., 2016) only performs pre-training and fine-tuning on Images. OSDA (Yang et al., 2021a) is an image sentiment classification model based on multi-view.

Multimodal model: MultiSentiNet (Xu and Mao, 2017) is a deep semantic network focusing on text-image sentiment analysis. HSN (Xu, 2017) is a hierarchical multimodal sentiment attention network based on image caption. MGNNS (Yang et al., 2021b) is a multi-channel graph neural networks with sentiment perception for multimodal sentiment detection. CLMLF (Li et al., 2022) uses multi-layer transformer for fusion and uses the idea of contrast learning for text-image sentiment detection. CIGNN (Wang et al., 2024) used attribute information to represent images, built two graph neural networks to model the global features of the dataset, and then carried out sentiment analysis.

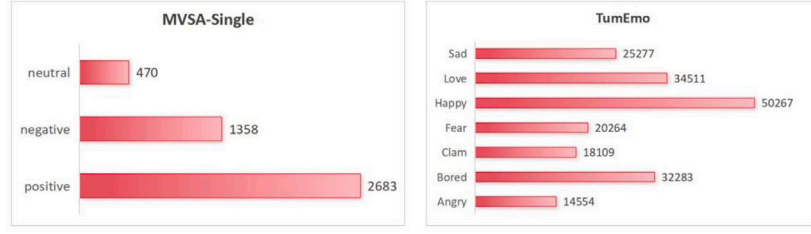


Fig. 3. Detailed statistical data of MVSA single and TumEmo datasets.

Table 2
Experimental results of different models on three datasets.

Modality	Model	MVSA-Single		TumEmo	
		ACC	F1	ACC	F1
Text	CNN (Kim, 2014)	0.6819	0.5590	0.6154	0.4774
	BiLSTM (Wang and Yang, 2020)	0.7012	0.6506	0.6188	0.5126
	BiACNN (Lai et al., 2015)	0.7036	0.6916	0.6212	0.5016
	TGNN (Huang et al., 2019)	0.7034	0.6594	0.6379	0.6362
Image	ResNet (He et al., 2016)	0.6467	0.6155	–	–
	OSDA (Yang et al., 2021a)	0.6675	0.6651	0.4770	0.3438
Image-Text	MultiSentiNet (Xu and Mao, 2017)	0.6984	0.6984	0.6309	0.5398
	HSAN (Xu, 2017)	0.6988	0.6690	0.6309	0.5398
	MGNNS (Yang et al., 2021b)	0.7377	0.7270	0.6672	0.6669
	CLMLF (Li et al., 2022)	<u>0.7533</u>	<u>0.7346</u>	–	–
	CIGNN (Wang et al., 2024)	0.7511	0.7333	<u>0.6738</u>	<u>0.6706</u>
	MAMSA(VIT-B/32)	0.7738	0.7768	0.6745	0.6723
	MAMSA(VIT-B/16)	0.7761	0.7684	0.6806	0.6792

4.4. Comparative experiments

To test the accuracy of the newly established model in multimodal sentiment recognition, we conduct validation on two public datasets and compared the model with unimodal and multimodal baseline models, and the experimental results are shown in Table 2. The results shown in bold are the results of the MAMSA model, and the underlined results are the best represented in the baseline model.

Quantitative analysis: from Table 2, it can be seen that compared with the TGNN, our model improved the experimental results by 7.29% and 10.9% on the MVSA-Single and by 4.27% and 4.7% on TumEmo. it can be seen that compared with OSDA, our model improved the experimental results by 10.86% and 10.33% on the MVSA-Single and by 20.36% and 33.54% on TumEmo. it can be seen that compared with the optimal baseline model, our model improved the experimental results by 2.28% and 3.38% on the MVSA-Single and by 0.68% and 0.86% on TumEmo. This indicates that the model can achieve the best performance in both three-category and seven-category sentiment analysis. MAMSA (VIT-B/32) and MAMSA (VIT-B/16) indicate that the models used in the feature encoding module are VIT-B/32 and VIT-B/16. Compared with VIT-B/32, VIT-B/16 has improved by 0.23% and 0.62% on the MVSA-Single and TumEmo datasets, respectively, indicating that VIT-B/16 performs better in feature extraction.

Qualitative analysis: Compared with single modal models, our model can capture sentiment expressions more comprehensively by integrating information from text and images, and has an advantage in handling complex emotional expressions. From Table 2, it is found that the image model is the worst for sentiment analysis, which is because the sentiment features in the image are too sparse that result to more noise, making it difficult for the model to acquire effective features for sentiment analysis. Compared with the baseline multimodal model, our model dynamically extracts common information between modalities, improving its ability to express features. In addition, sentiment attention can make text and images focus on their own sentiment features, improving the accuracy of sentiment analysis. Finally, hierarchical fusion combines contextual information to understand sentiment expression, such as inferring the user's true sentiment through the context in the text and the scene in the image.

Table 3
Experimental results of different models on three datasets.

Model	MVSA-Single		TumEmo	
	ACC	F1	ACC	F1
Text-image	0.6919	0.6910	0.5930	0.5796
+AAI	0.7627	0.7593	0.6333	0.6295
+SA	0.7694	0.7675	0.6731	0.6710
+HF	0.7761	0.7684	0.6808	0.6792

“A2B-C” indicates that the graph constructed on the “A” dataset is used for the “B” dataset, where C ∈ {text, image}, A ∈ {MVSA-Multiple, MVSA-Single, TumEmo}, and B ∈ {MVSA-Multiple, MVSA-Single, TumEmo}. For example, “S2T-image” indicates that the image graph constructed on the MVSA-Single dataset is used for the TumEmo dataset.

4.5. Ablation experiment

In order to further analyze the influence of different components of the model on the performance of the whole model, we designed four sets of ablation experiments on MVSA-Single and-TumEmo datasets. The experimental results are shown in Table 3.

As we gradually introduce the proposed modules, we can clearly observe their significant contribution to the experimental results, which further validates the effectiveness and necessity of each module design. Compared with Text-image, the experimental results of +AAI improved by 7.08% and 4.03% on the two datasets, respectively. Specifically, the adaptive attention interaction module (+AAI), through its dynamically adjustable features, can flexibly capture the most critical information segments for the current task from early fusion features, improving the performance of the model. Subsequently, based on AAI, we further added an sentiment embedding matrix to enhance the sentiment attention mechanism (+SA). Compared to using only AAI, the experimental results of +SA improved by 0.67% and 3.98% on both datasets, respectively. This improvement not only demonstrates the effectiveness of the sentiment embedding matrix, but also reveals how it assists attention mechanisms in more accurately focusing on text words and image regions that can deeply express sentiment. The introduction of sentiment embedding matrix makes the model more sensitive and



id	image	text	label	θ
484		#frozen #cold #canada #winnipeg #natural #photography #art #snow #ice #deserted #shed	negative	0.308
3970		"@tuscanigram: . Are you on the beach? ?We aren' t ... But Pienza is always zippy! ? Touris...	positive	0.743
4125		RT @TransferSources: The Chelsea fans who were racist to a black man in the Paris Metro have been banned from football matches for 5 years ...	negative	0.583
4432		Why are you feeling dismal? Take the quiz: http://t.co/UEwYjGvMM2 http://t.co/M8IU5aazXT	positive	0.678

Fig. 4. Text-image Pairs with learned θ .Table 4
Order of hierarchical fusion.

Model	MVSA-Single		TumEmo	
	ACC	F1	ACC	F1
(1) text-image-cat	0.7761	0.7684	0.6808	0.6792
(2) image-text-cat	0.7716	0.7632	0.6758	0.6736
(3) cat-text-image	0.5942	0.5900	0.4637	0.4553
(4) cat-image-text	0.5904	0.5813	0.5890	0.5807
(5) text-cat-image	0.5943	0.5896	0.4673	0.4558
(6) image-cat-text	0.5923	0.5879	0.4677	0.4426

Note: Text represents sentiment-text association expression, image represents sentiment-image association expression, and cat represents multimodal common information. X-Y-Z represents X is the query Q for the first layer of Attention, Y represents the key K and value V for the first layer of Attention, the result of their interaction is the query Q for the second layer of Attention, Z represents the key K and value V for the second layer of Attention, where $X \in \text{text}, \text{image}, \text{cat}$, $Y \in \text{text}, \text{image}, \text{cat}$, $Z \in \text{text}, \text{image}, \text{cat}$.

in-depth in understanding sentiment content, thereby improving the overall performance of sentiment analysis. Finally, in order to maximize the complementarity between multiple modalities, we added a hierarchical fusion module (+HF) on top of +SA. The experimental results show that the addition of +HF further improves the performance of the model on both datasets, which fully demonstrates the effectiveness of the hierarchical fusion strategy in enhancing the performance of multimodal sentiment analysis tasks.

Through this series of module addition and validation processes, we not only demonstrated the independent contribution of each module to improving model performance, but also revealed how they work together to build a more efficient and powerful multimodal sentiment analysis model.

4.6. Hierarchical fusion experiment

In order to analyze the influence of fusion order in the hierarchical fusion module on model performance, we designed six sets of experiments with different fusion orders. The dataset used are still MVSA-Single and TumEmo, and the experimental results are shown in Table 4.

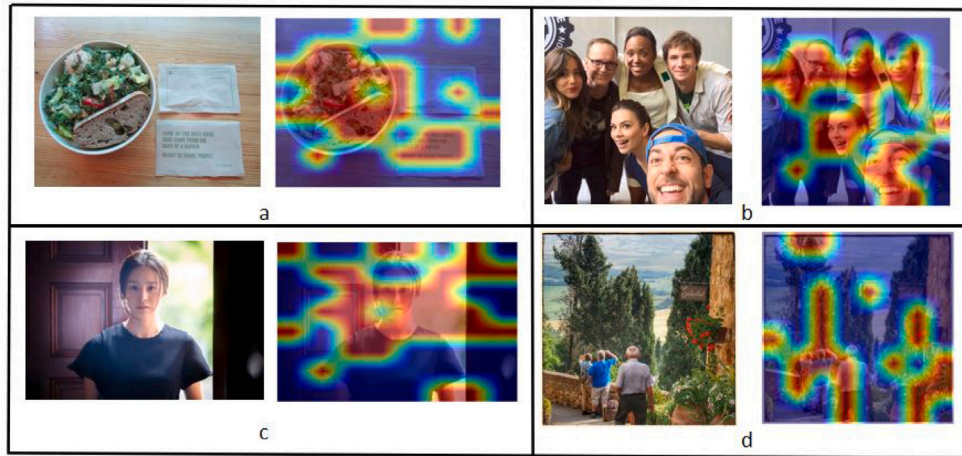
From Table 4, it is found that changing the order of fusion will result in poorer performance. Compared to (2), (1) performs better because text plays a dominant role in multimodal fusion so that it better focus on the image regions that are relevant to the text. In the first layer of Attention, text is dominant, and the attention mechanism can focus on the text-related areas of the image, thereby improving the performance of the model. Compared with (1) and (2), the performance of (3) - (6) is significantly reduced. This is because *cat* is in an unreasonable position. For example, in *cat-text-image*, in the first layer of Attention, Q is multimodal common information representation, while K and V are sentiment-text associations representation. This setting will result in the loss of sentiment information in the sentiment-text association representation, ultimately leading to a decrease in the performance of the model. In summary, a reasonable fusion sequence will improve the performance of the model.

4.7. Adaptive attention interaction analysis

As shown in formula (4), the scalar θ learned by AAI indicates the contribution of text and images during fusion. Intuitively, the value of θ may affect the performance of the MAMSA model. Therefore, in order to verify the effectiveness of the designed adaptive attention interaction, we compared AAI with different predefined θ and analyzes the learned θ of four sets of text-image pairs. The results of the predefined θ and MAMSA are shown in Table 5.

The ten experiments we designed set the predefined θ parameter as a series of nine different values that gradually increased from 0.1 to 0.9. This detailed division can comprehensively examine the specific impact of changes in θ values on the results of sentiment analysis. As shown in Table 5, each unique θ value corresponds to different sentiment analysis results. In addition, compared to the predefined θ , the adaptive θ value learned by the model performs better.

It is particularly noteworthy that compared to these static and predefined θ values, the adaptive θ values obtained by the model through the adaptive attention mechanism exhibit superior performance. This conclusion is strongly supported by Fig. 4, which clearly depicts the correspondence between different text-image pairs and their learned θ values. This one-to-one relationship deeply reveals that the

Fig. 5. Text-image Pairs with learned θ .Table 5
Results of different θ values.

Model	ACC	F1
$\theta = 0.9$	0.7690	0.7671
$\theta = 0.8$	0.7650	0.7538
$\theta = 0.7$	0.7672	0.7605
$\theta = 0.6$	0.7750	0.7675
$\theta = 0.5$	0.7684	0.7599
$\theta = 0.4$	0.7737	0.7796
$\theta = 0.3$	0.7670	0.7600
$\theta = 0.2$	0.7605	0.7494
$\theta = 0.1$	0.7517	0.7472
MAMSA	0.7761	0.7684

AAI (Adaptive Attention Interaction) module can adaptively learn the optimal θ value for each text-image pair. Specifically, for the text-image pair with id 484, the model learns a θ value of 0.308; On the contrary, the text-image pair with id 3970 learns a θ value of 0.743. This adaptive learning approach breaks the limitations of fixed weight settings in traditional methods, allowing the model to more flexibly respond to diverse sentiment expression scenarios. If relying solely on predefined θ values (such as $\theta = 0.5$), although it can provide a basic fusion framework, it often cannot accurately capture the unique sentiment features of each text-image pair, resulting in suboptimal analysis results.

4.8. Sentiment visualization experiment

In order to deeply analyze whether the attention mechanism can effectively focus on the image region that best expresses sentiments under the guidance of the sentiment embedding matrix. We visualize the attention weights of SIA, and the results of attention visualization are shown in Fig. 5.

From Fig. 5, it can be observed that under the guidance of the sentiment embedding matrix, the model accurately focuses on the regions that have a decisive impact on sentiment expression and assigns higher attention weights to these regions accordingly. For example, in Fig. 5a, the model focuses more attention on the two key elements of food and paper with words. Food, as one of the core elements in images, often expresses the sentiment of bloggers; And the text on the paper with words is likely to directly reveal the sentiment tendency of the image. In Fig. 5b, the smiling face of the character is undoubtedly the most direct and powerful way to convey positive sentiment. The model successfully locked onto this critical region through attention mechanism and gave it high attention. This precise sentiment localization

ability not only demonstrates the deep foundation of the model in the field of image sentiment analysis, but also provides strong support for its wider application in multimodal sentiment analysis.

5. Discussion on potential extensions in video sentiment analysis

A video typically conveys a same sentiment through multimodal information such as text, sound, and visual images. The key challenge that currently needs to be tackled is how to fuse and maximize the utilization of information between different heterogeneous data within the same video. Our proposed model, MAMSA, is capable of extracting common information across different modalities, enhancing the performance of feature expression in each modality, and efficiently integrating multimodal heterogeneous data including text, sound, and visual images in videos. It minimizes the differences between modalities as much as possible, ultimately enabling the recognition and analysis of emotions in videos. However, our model did not take into account the temporal information, so there are certain limitations in predicting video sentiment.

6. Conclusion

To address the problem that extracting a fixed amount of common information does not apply to all posts, we propose a multimodal sentiment analysis model based on multiple attention. Compared with previous works, our proposed MAMSA is able to dynamically extract common information between modalities, effectively improving the information utilization and feature expression capabilities of multimodal data. Meanwhile, our method enables the model to focus on learning those features that are crucial to sentiment, significantly enhancing the accuracy of sentiment recognition.

Our method provides a new perspective and approach for multimodal information processing, but it does not consider temporal information, which limits its application in the field of video sentiment analysis. In the future, we will supplement our research from two aspects: (1) we will consider using SMOTE and class weighting to handle imbalanced datasets; (2) We will consider temporal characteristics in the model to enhance its generalization ability; (3) We will introduce external knowledge bases related to sentiment, such as sentiment dictionaries and social media trends, to enable more precise extraction of sentiment-related features.

CRedit authorship contribution statement

Hongbin Wang: Writing – review & editing, Methodology. **Chun Ren:** Writing – original draft, Software, Methodology. **Zhengtao Yu:** Writing – review & editing.

Ethical and informed consent for data used

The authors use an open source dataset, which does not have ethical issues. The dataset is sourced from <https://twitter.com> and <http://tumblr.com>.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

Acknowledgments

This work was supported by the Nation Natural Science Foundation of China under Grant 61966020, the Yunnan Natural Science Funds under Grant 202201AT070157.

Data availability

The authors use an open source dataset, the dataset is sourced from <https://twitter.com> and <http://tumblr.com>.

References

- Chen, D., Su, W., Wu, P., Hua, B., 2023. Joint multimodal sentiment analysis based on information relevance. *Inf. Process. Manage.* 60 (2), <http://dx.doi.org/10.1016/j.ipm.2022.103193>.
- Hazarika, D., Zimmermann, R., Poria, S., 2020. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In: *Proceedings of the 28th ACM International Conference on Multimedia. MM '20*, Association for Computing Machinery, New York, NY, USA, pp. 1122–1131. <http://dx.doi.org/10.1145/3394171.3413678>.
- He, J., Yanga, H., Zhang, C., Chen, H., Xua, Y., Hu, Z., 2022. Dynamic invariant-specific representation fusion network for multimodal sentiment analysis. *Intell. Neurosci.* 2022, <http://dx.doi.org/10.1155/2022/2105593>.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR*, pp. 770–778. <http://dx.doi.org/10.1109/CVPR.2016.90>.
- Huang, L., Ma, D., Li, S., Zhang, X., Wang, H., 2019. Text level graph neural network for text classification. In: Inui, K., Jiang, J., Ng, V., Wan, X. (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. EMNLP-IJCNLP*, Association for Computational Linguistics, Hong Kong, China, pp. 3444–3450. <http://dx.doi.org/10.18653/v1/D19-1345>.
- Huang, C., Zhang, J., Wu, X., Wang, Y., Li, M., Huang, X., 2023. TeFNA: Text-centered fusion network with crossmodal attention for multimodal sentiment analysis. *Knowl.-Based Syst.* 269, 110502. <http://dx.doi.org/10.1016/j.knsys.2023.110502>.
- Kim, Y., 2014. Convolutional neural networks for sentence classification. In: Moschitti, A., Pang, B., Daelemans, W. (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. EMNLP*, Association for Computational Linguistics, Doha, Qatar, pp. 1746–1751. <http://dx.doi.org/10.3115/v1/D14-1181>.
- Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. <http://dx.doi.org/10.48550/arXiv.1412.6980>, CoRR abs/1412.6980.
- Lai, S., Xu, L., Liu, K., Zhao, J., 2015. Recurrent convolutional neural networks for text classification. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. AAAI '15*, AAAI Press, pp. 2267–2273.
- Le, H.-D., Lee, G.-S., Kim, S.-H., Kim, S., Yang, H.-J., 2023. Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning. *IEEE Access* 11, 14742–14751. <http://dx.doi.org/10.1109/ACCESS.2023.3244390>.
- Li, Z., Xu, B., Zhu, C., Zhao, T., 2022. CLMLF: A contrastive learning and multi-layer fusion method for multimodal sentiment detection findings of the association for computational linguistics: NAACL 2022. pp. 2282–2294. <http://dx.doi.org/10.18653/v1/2022.findings-naacl.175>.
- Lin, F., Liu, S., Zhang, C., Fan, J., Wu, Z., 2023. StyleBERT: Text-audio sentiment analysis with Bi-directional style enhancement. *Inf. Syst.* 114, 102147. <http://dx.doi.org/10.1016/j.is.2022.102147>.
- Liu, H., Li, K., Fan, J., Yan, C., Qin, T., Zheng, Q., 2023a. Social image-text sentiment classification with cross-modal consistency and knowledge distillation. *IEEE Trans. Affect. Comput.* 14 (4), 3332–3344. <http://dx.doi.org/10.1109/TAFFC.2022.3220762>.
- Liu, X., Xu, Z., Huang, K., 2023b. Multimodal emotion recognition based on cascaded multichannel and hierarchical fusion. *Comput. Intell. Neurosci.* <http://dx.doi.org/10.1155/2023/9645611>.
- Niu, T., Zhu, S., Pang, L., El Saddik, A., 2016. Sentiment analysis on multi-view social data. In: Tian, Q., Sebe, N., Qi, G.-J., Huet, B., Hong, R., Liu, X. (Eds.), *MultiMedia Modeling. Springer International Publishing, Cham*, pp. 15–27.
- Quan, Z., Sun, T., Su, M., Wei, J., Zhang, X., Zhong, S., 2022. Multimodal sentiment analysis based on nonverbal representation optimization network and contrastive interaction learning. In: *2022 IEEE International Conference on Systems, Man, and Cybernetics. SMC*, pp. 3086–3091. <http://dx.doi.org/10.1109/SMC53654.2022.9945514>.
- Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (Eds.), *Proceedings of the 38th International Conference on Machine Learning. In: Proceedings of Machine Learning Research*, vol. 139, PMLR, pp. 8748–8763.
- Sun, H., Liu, J., Chen, Y.-W., Lin, L., 2023. Modality-invariant temporal representation learning for multimodal sentiment classification. *Inf. Fusion* 91 (C), 504–514. <http://dx.doi.org/10.1016/j.inffus.2022.10.031>.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS '17*, Curran Associates Inc., Red Hook, NY, USA, pp. 6000–6010.
- Wang, H., Ren, C., Yu, Z., 2024. Multimodal sentiment analysis based on cross-instance graph neural networks. *Appl. Intell.* 3403–3416. <http://dx.doi.org/10.1007/s10489-024-05309-0>.
- Wang, Z., Yang, B., 2020. Attention-based bidirectional long short-term memory networks for relation classification using knowledge distillation from BERT. In: *2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress. DASC/PiCom/CBDCCom/CyberSciTech*, pp. 562–568. <http://dx.doi.org/10.1109/DASC-PiCom-CBDCCom-CyberSciTech49142.2020.00100>.
- Wu, K., Peng, H., Chen, M., Fu, J., Chao, H., 2021. Rethinking and improving relative position encoding for vision transformer. In: *2021 IEEE/CVF International Conference on Computer Vision. ICCV*, pp. 10013–10021. <http://dx.doi.org/10.1109/ICCV48922.2021.00988>.
- Xiao, L., Wu, X., Wu, W., Yang, J., He, L., 2022. Multi-channel attentive graph convolutional network with sentiment fusion for multimodal sentiment analysis. In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP*, pp. 4578–4582. <http://dx.doi.org/10.1109/ICASSP43922.2022.9747542>.
- Xiao, L., Wu, X., Yang, S., Xu, J., Zhou, J., He, L., 2023. Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis. *Inf. Process. Manage.* 60 (6), 103508. <http://dx.doi.org/10.1016/j.ipm.2023.103508>.
- Xu, N., 2017. Analyzing multimodal public sentiment based on hierarchical semantic attentional network. In: *2017 IEEE International Conference on Intelligence and Security Informatics. ISI*, pp. 152–154. <http://dx.doi.org/10.1109/ISI.2017.8004895>.
- Xu, M., Liang, F., Su, X., Fang, C., 2022. CMJRT: Cross-modal joint representation transformer for multimodal sentiment analysis. *IEEE Access* 10, 131671–131679. <http://dx.doi.org/10.1109/ACCESS.2022.3219200>.
- Xu, N., Mao, W., 2017. MultiSentNet: A deep semantic network for multimodal sentiment analysis. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. CIKM '17*, Association for Computing Machinery, New York, NY, USA, pp. 2399–2402. <http://dx.doi.org/10.1145/3132847.3133142>.
- Yang, X., Feng, S., Wang, D., Zhang, Y., 2021a. Image-text multimodal emotion classification via multi-view attentional network. *IEEE Trans. Multimed.* 23, 4014–4026. <http://dx.doi.org/10.1109/TMM.2020.3035277>.
- Yang, X., Feng, S., Zhang, Y., Wang, D., 2021b. Multimodal sentiment detection based on multi-channel graph neural networks. In: *Annual Meeting of the Association for Computational Linguistics. Vol. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. vol. 1*, Long Papers.
- Yang, D., Huang, S., Kuang, H., Du, Y., Zhang, L., 2022. Disentangled representation learning for multimodal emotion recognition. In: *Proceedings of the 30th ACM International Conference on Multimedia. MM '22*, Association for Computing Machinery, New York, NY, USA, pp. 1642–1651. <http://dx.doi.org/10.1145/3503161.3547754>.
- Zeng, D., Chen, X., Song, Z., Xue, Y., Cai, Q., 2023. Multimodal interaction and fused graph convolution network for sentiment classification of online reviews. *Mathematics* 11 (10), <http://dx.doi.org/10.3390/math11102335>.
- Zhang, Z., Wang, Z., Li, X., Liu, N., Guo, B., Yu, Z., 2021. ModalNet: an aspect-level sentiment classification model by exploring multimodal data with fusion discriminant attentional network. 24 (6), 1957–1974. <http://dx.doi.org/10.1007/s11280-021-00955-7>.
- Zhao, X., Chen, Y., Chen, Y., Liu, S., Tang, B., 2022. HMAI-BERT: Hierarchical multimodal alignment and interaction network-enhanced BERT for multimodal sentiment analysis. In: *2022 IEEE International Conference on Multimedia and Expo. ICME*, pp. 1–6. <http://dx.doi.org/10.1109/ICME52920.2022.9859747>.

- Zhao, X., Chen, Y., Liu, S., Tang, B., 2023. Shared-private memory networks for multimodal sentiment analysis. *IEEE Trans. Affect. Comput.* 14 (4), 2889–2900. <http://dx.doi.org/10.1109/TAFFC.2022.3222023>.
- Zhou, R., Guo, W., Liu, X., Yu, S., Zhang, Y., Yuan, X., 2023a. AoM: Detecting aspect-oriented information for multimodal aspect-based sentiment analysis. In: *Findings of the Association for Computational Linguistics: ACL 2023*. pp. 8184–8196. <http://dx.doi.org/10.18653/v1/2023.findings-acl.519>.
- Zhou, S., Wu, X., Jiang, F., Huang, Q., Huang, C., 2023b. Emotion recognition from large-scale video clips with cross-attention and hybrid feature weighting neural networks. *Int. J. Environ. Res. Public Health* 20 (2), <http://dx.doi.org/10.3390/ijerph20021400>.
- Zhu, C., Chen, M., Zhang, S., Sun, C., Liang, H., Liu, Y., Chen, J., 2023. SKEAFN: Sentiment knowledge enhanced attention fusion network for multimodal sentiment analysis. *Inf. Fusion* 100 (C), <http://dx.doi.org/10.1016/j.inffus.2023.101958>.