International Journal of Applied Mathematics

Volume 38 No. 1s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

HIERARCHICAL ATTENTION MECHANISMS FOR MULTIMODAL SENTIMENT ANALYSIS IN USER-GENERATED SOCIAL MEDIA CONTENT

Yuanyan Li[1,2], Intan soliha Ibrahim[1], Soon Fook Fong[3]

[1]Faculty of Social Sciences and Humanities, University Malaysia Sabah,
Kota Kinabalu, 88400, Sabah, Malaysia
[2]Department of Design and Communication, College of Modern Economics
and Management Jiangxi University of Finance and Economics, Gongqing
City 332020, Jiangxi, China
[3]Academy of Arts and Creative Technology, University Malaysia Sabah，
Kota Kinabalu, 88400，Sabah ，Malaysia
E-mail: mypaper2023@163.com, Resr.Intan@outlook.com,
kikilee1221@sina.com
Oricd: 0000-0001-8585-6008, https://orcid.org/0000-0002-8483-4824,
https://orcid.org/0009-0009-0393-4777

Abstract

In multimodal sentiment analysis, effective extraction of fused multimodal features and the use of attention mechanism to improve analysis accuracy are key issues. The research constructs multi-source heterogeneous data collection architecture to collect user-generated content containing text, images, and videos from social media, short videos, news media, and other platforms. This study proposes a multimodal sentiment analysis model based on bidirectional long short-term memory network and hierarchical attention mechanism. It introduces the sentiment attention capsule structure and combines it with a sentiment loss function to guide the generation of specific sentiment responses. The experiments are based on a multimodal emotion intensity dataset and a multimodal dialog action dataset, which are compared with multiple models. The results showed that the model had an accuracy of 90.285%, precision of 90.442%, recall of 89.517%, F1-score of 89.980%, AUC value of 0.935, and training time of 9.56 hours. Moreover, the overall performance was significantly better with 64.887 ms inference speed, 2.487 GB memory usage, and 94.209% training stability. The study offers a fresh approach to raising sentiment analysis accuracy, and the model's attention mechanism and data fusion architecture are novel. It has application value for intelligent dialogue, emotion monitoring and other scenarios, and provides reference for model optimization and technology landing in this field.

Received: July 20, 2025

Keywords: Multimodal sentiment analysis; Hierarchical attention mechanism; Sentiment attention capsule; Multi-source heterogeneous data collection; Model performance evaluation.

1.Introduction

User-generated content (UGC) is growing at an explosive rate in social media, short video platforms, news media and other online ecologies. Multimodal data such as text, image, video, etc. together constitute a complex carrier of sentiment expression [1-2]. The trend of sentiment analysis (SA) from unimodal to multimodal has received attention from scholars [3]. Das R et al. explored the trend of SA from unimodal to multimodal development and related contents. The study conducted a comprehensive research on different SA methods, applications, challenges, and resources, focusing on the consideration of audio, video, and other channel data in multimodal SA. The results indicated that multimodal SA enabled more depth and accuracy in emotion detection by fusing multiple data streams [4]. Singh U et al. addressed the challenges posed by the growth in data size, subjectivity, and diversity to enhance the efficiency of existing SA techniques by conducting a comprehensive study of various literatures dealing with different aspects of SA. The article identified these unresolved challenges to clarify the future prospects of multimodal SA and provide guidelines to researchers [5]. Ghorbanali et al. carried out a thorough comparative analysis of signal analysis techniques, difficulties, uses, and patterns, paying particular attention to deep learning-based techniques for multimodal SA. The issues with multimodal SA that were addressed in this work included missing data, modal heterogeneity, fusion techniques, intermodal interactions, irrelevance, and redundant or insufficient data information. By analyzing the shortcomings of current research, outlining potential future remedies, and evaluating current difficulties, the study assessed the methodologies' future trajectory [6]. Sun L et al. proposed efficient multimodal Transformer with dual-level feature restoration (EMT-DLFR) to address the two major challenges of inefficient cross-modal interaction and missing random modal features in multimodal SA. EMT avoided the secondary extension cost of earlier approaches by using the statement-level representation of each modality as the global context for interacting with local features. To draw in high-level representations of both whole and incomplete input, DLFR used twin learning and low-level feature reconstruction. Experiments proved that the method performed better in both complete and partial modal contexts [7]. However, traditional studies face multiple challenges in dealing with multimodal data [8]. First, feature fusion of heterogeneous data from multiple sources is not efficient enough. Early methods often adopt simple splicing or fixed weight fusion strategies, which are difficult to dynamically capture the complementarity and emotional relevance between different modalities. Second, the attention mechanism (AM) lacks

Received: July 20, 2025

hierarchical design. The traditional single-level attention model cannot effectively distinguish the econd nal primary econddary relationships in the contextual information, leading to the dilution of key emotional features by noise. Third, the refinement of emotion modeling is insufficient. Most models do not explicitly introduce emotional polarity constraints, making it difficult to generate response content that meets specific emotional orientations. Fourth, there are limitations at the data collection level. Traditional architectures do not support the breadth of coverage and modal diversity of multi-platform data sufficiently, leading to the representative bias of training data [9-11]. To address the above problems, the research constructs multi-source heterogeneous data collection architecture to integrate multimodal data from social media, short video platforms and news media. The innovative aspect of the research is the realization of cross-platform data collection through the use of the Octoparse crawler, the MitmProxy packet grabbing tool, and the Requests library. The study designs a hierarchical AM that dynamically captures emotional focus and contextual dependencies in text sequences through two-layer attention modeling at the word and sentence levels. The study also introduces a sentiment attention capsule structure that preserves differences in the activation levels of multimodal features. Finally, a sentiment loss function is embedded in model training to couple the sentiment classification task with the sequence generation task. This guides the model to generate reply content with explicit sentiment tendencies. This research is expected to overcome the technical limitations of traditional methods in feature fusion, attention allocation, and emotional constraints to provide a more robust solution for multimodal SA.

2. Hierarchical Ams and modeling of multimodal Sas

2.1 Multimodal feature extraction and fusion

To enhance the accuracy and comprehensiveness of multimodal SA, the study adopted the strategy of collecting data from multiple online platforms. These platforms provide rich UGC, including text, images, videos, and other types of data, and are able to provide diverse input information for the SA task [12]. Therefore, the study constructs a multi-source heterogeneous data collection architecture to ensure that representative data can be collected from different platforms. The multi-platform UGC data collection architecture is shown in Figure 1.

Received: July 20, 2025

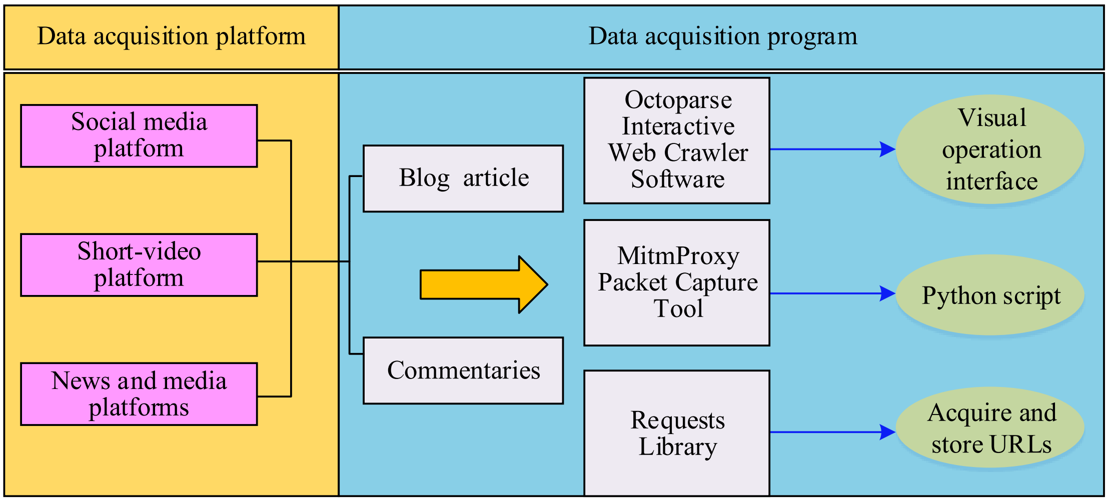


Figure 1 Multi-platform UGC data collection architecture

Figure 1 displays the architecture of the data collection platform and data collection program. In the data collection platform section, it contains social media platforms, short video platforms, and news media platforms. These platforms generate blog posts and comments. These contents point to the data collection program through arrows. The program utilizes three tools for data collection. Octoparse interactive web crawler software output corresponds to the visualization interface. MitmProxy packet-crawling tool output associated Python scripts. Requests library is used to fetch and store URLs. Social media UGC is associated with communities and topics as shown in Figure 2.

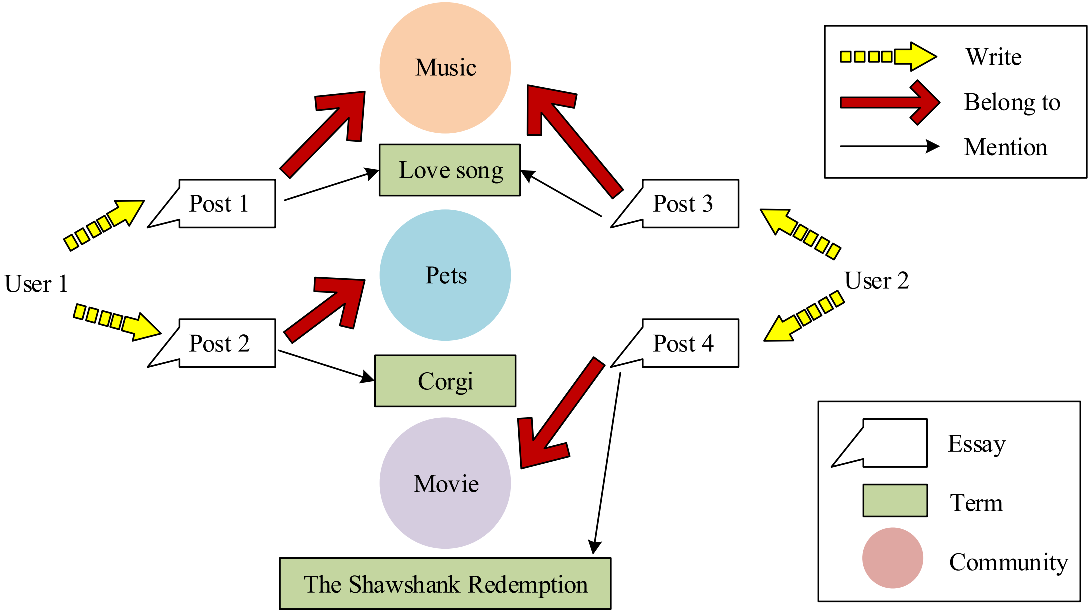


Figure 2 Social media UGC associated with communities and themes

Figure 2 shows the association of UGC with communities of interest and topics. User 1 and User 2 generate post 1, post 2, post 3, and post 4 by writing. These posts belong to the

Received: July 20, 2025

association with communities of interest along with mentions pointing to subject terms. The network architecture for handling user-generated social media content is shown in Figure 3.

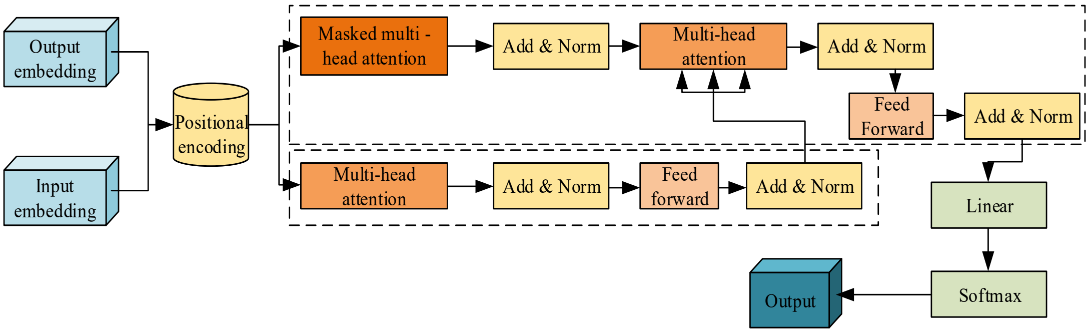


Figure 3 Network architecture for processing user-generated social media content

Figure 3 illustrates a network architecture for processing user-generated social media content, closely related to feature extraction and fusion in multimodal SA. The input embeddings and output embeddings are positionally encoded and then enter the core processing modules. First is the masked multimodal attention module. This module masks irrelevant parts when processing sequence data, ensuring that the model focuses on valid information. Subsequently, the data distribution is stabilized by summing and normalization operations to enhance the training effect. Next, the multi-head AM captures features from multiple dimensions to enhance the understanding and integration of multimodal information. After summation and normalization again, the data enters the feed-forward network to further refine the features. Finally, after linear transformation and Softmax function, the processing results are transformed into the final output to complete tasks such as sentiment classification. The study utilizes the BERT model to classify the sentiment of users' social media and obtain the sentiment distribution of intelligent replies. The language robot is guided to provide verbal responses with particular emotional traits by combining the emotion loss function. Enhancing the dialog impact and expanding the model's applicability are two benefits of adding the emotion loss function. $Tokn$ denotes image region features. The input sequence is processed by BERT to generate a multimodal feature vector sequence $W=\{w_1,w_2,\cdots,w_n\}$, where are the cross-modal fused features [13]. Moreover, w_i is used as the input sequence of bidirectional long short-term memory network (Bi-LSTM) as shown in Equation (1).

$$h_i = BiLSTM\left(h_{i-1}, w_i\right) \quad (1)$$

In Equation (1), h_i and h_{i-1} denote the context-aware features of text words i and $i-1$. The weight coefficient λ_i is calculated as shown in Equation (2).

$$\begin{cases} \lambda_i = \dfrac{\exp\left(M_i\right)}{\sum_{j=1}^{n}\exp\left(M_i\right)} \\ M_i = \tanh\left(W_m \cdot h_i\right) \end{cases} \quad (2)$$

In Equation (2), W_m denotes the within-modality attentional weight matrix. M_i

denotes the word-level emotional activation value. Weight λ_i denotes the attentional weight of word i. By attentional weighting, the word-level affective information is aggregated, and the image modality is obtained in the same way. Summing the implicit layer vectors weighted by the attentional weights, the local sentiment feature vector $cVec$ of the text modality is obtained, as shown in Equation (3).

$$cVec = \sum_{i=1}^{n} \lambda_j \cdot h_i \quad (3)$$

The global multimodal emotion representation is obtained by calculating the inter-modal attention weights for each local representation of each modality, then summing and weighting them. The emotion probability vector $eVec$ is calculated as shown in Equation (4).

$$eVec = Softmax\left(Dense\left(c\right)\right) \quad (4)$$

The probability distribution of each sentiment category is produced by the Softmax layer after the fully connected layer (FCL) maps the global representations to the sentiment category space. The prediction response sequence is set to be $Y = \left(y_1, y_2, \cdots, y_{T_y}\right)$, while the real sentiment category is $Y = \left(y_1, y_2, \cdots, y_{T_y}\right)$ [14]. The sequence-to-sequence model (Seq2Seq) is fused to construct a hierarchical multi-loss function for user-generated social media multimodal content (text, image, video). The sequence generation, sentiment classification, and sentiment polarization constraints are fused to optimize the multimodal SA with response generation as shown in Equation (5).

$$L_{total}\left(\theta\right) = \lambda_1 \cdot L_{s2s}\left(\theta\right) + \lambda_2 \cdot L_{EC}\left(\theta\right) + \lambda_3 \cdot L_{EP}\left(\theta\right) \quad (5)$$

In Equation (5), $L_{total}\left(\theta\right)$ is used to integrate multi-task optimization. θ is the model parameters, including hierarchical attention weights, multimodal encoder parameters, etc. $L_{s2s}\left(\theta\right)$ is the multimodal sequence loss. $L_{EC}\left(\theta\right)$ is the multimodal sentiment categorization loss for optimizing sentiment category prediction based on fused features of text, image, and video. $L_{EP}\left(\theta\right)$ is sentiment polarization loss, which is used to enhance the response sentiment intensity and improve the social interaction infectiousness. λ_1, λ_2, and λ_3 are loss weights, balanced multitasking, and sum to 1. The $L_{s2s}\left(\theta\right)$ function is shown in Equation (6).

$$L_{s2s}\left(\theta\right) = -\log\left(Y \middle| X\right) = -\sum_{i=1}^{T_y} \log p\left(y_i \middle| y_1, y_2, \cdots, y_{i-1}, X\right) \quad (6)$$

In Equation (6), y_i is the uniquely hot coding of the real sentiment category. T denotes the length of the reply sequence, which ensures that the reply is semantically related to the input and fits the contextual logic of the social media text. $eVec$ is the sentiment vector after BERT encodes the user content. The cross-entropy loss enables the model to accurately capture the sentiment tendency of user content, as shown in Equation (7).

$$L_{EC}\left(\theta\right) = -\sum_{i=1}^{m} y_i \log\left(eVec\right) \quad (7)$$

In Equation (7), m denotes the number of emotion categories. The emotional impact of

Received: July 20, 2025

social media interactions is enhanced by minimizing the loss and increasing the density of emotional words in replies. The formula for $L_{EP}(\theta)$ is shown in Equation (8).

$$L_{EP}(\theta) = -\frac{1}{T_y}\sqrt{\sum_{i=1}^{T_y}\sum_{j=1}^{25}\left(yEmo_{ij}\right)^2} \quad (8)$$

In Equation (8), T_y means the number of emotion polarity words in the response. $yEmo_{ij}$ means the j th element of the i th sentiment word vector in the sequence.

2.2 Hierarchical AM with contextual modeling

The focus will be on using the hierarchical AM with contextual modeling on this foundation to further enhance the accuracy and robustness of SA models, following an exploration of how to extract rich features from multimodal data and fuse them into a coherent representation [15]. The hierarchical AM with contextual modeling in multimodal SA is shown in Figure 4.

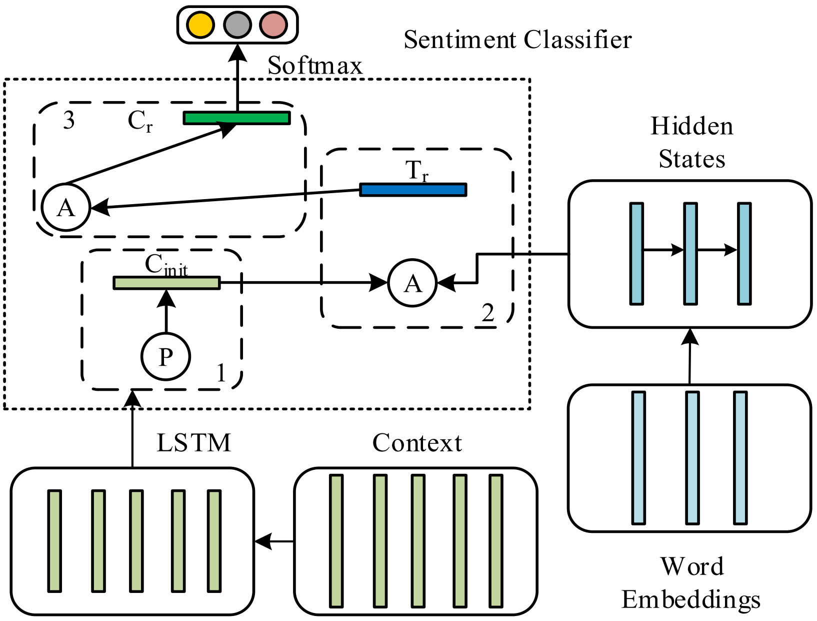


Figure 4 Hierarchical AM in multimodal SA with contextual modeling

Figure 4 illustrates the hierarchical AM with context modeling process for a multimodal sentiment analysis model (MSAM). In this model, firstly, the input text data will be represented by the word embedding layer, and then passed to LSTM for sequence modeling to generate hidden states. On this basis, the model weights the contextual information through hierarchical AM to capture the important parts of the text. Specifically, the hierarchical AM in Figure consists of two parts. The context-based attention weight computation, which may concentrate on the sentiment data at various context time steps, is the initial component. Contextual modeling, which is used in the second section to further process the sequence data,

Received: July 20, 2025

allows the model to incorporate the text's contextual information to increase sentiment classification accuracy. Finally, the model performs sentiment classification through the Softmax layer, mapping the input data to different sentiment categories.

In the multimodal SA task, the model must efficiently focus on important emotional information and suppress irrelevant noise. To address this, a new AM structure called the sentiment attention capsule is proposed [16]. The introduction of the sentiment attention capsule into the sentiment classification model enables this hybrid model to dynamically generate vectors adapted to different sentiment expressions within a limited representation space. This allows the model to capture sentiment signals more accurately. The structure of sentiment attention capsule is shown in Figure 5.

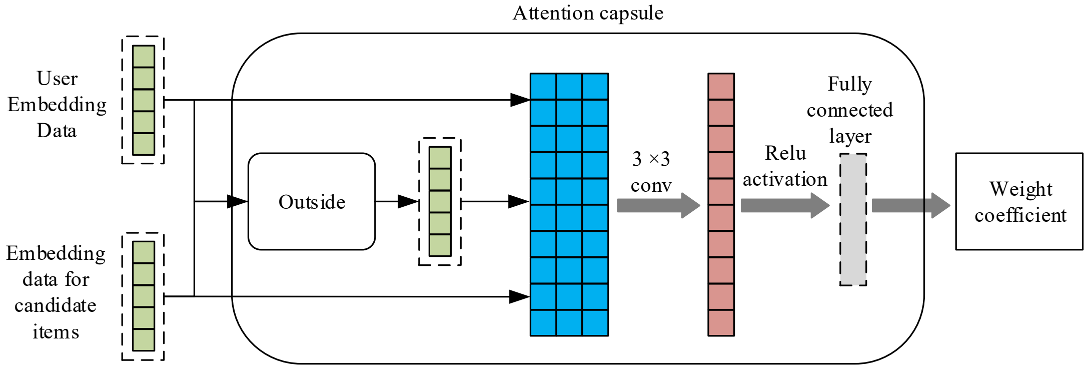


Figure 5 Structure of sentiment attention capsule

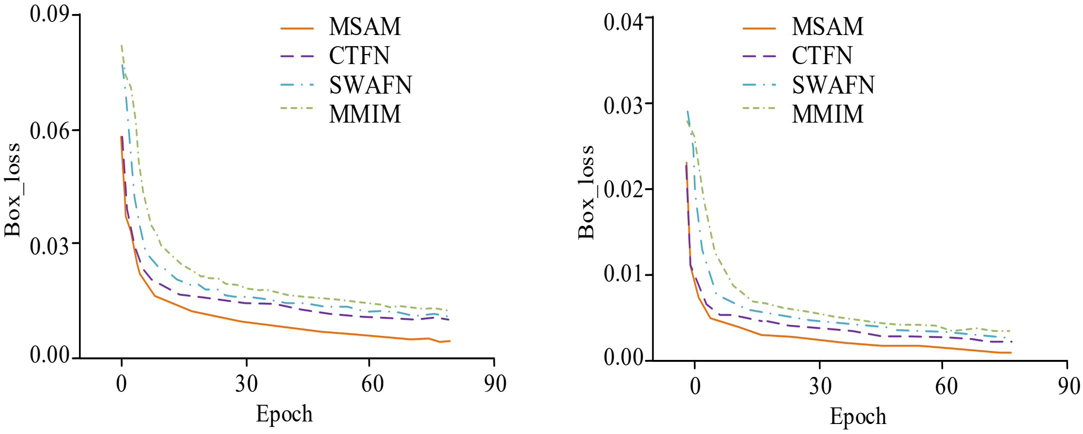
Based on the sentiment attention capsule structure, Figure 5 displays how hierarchical AM is applied in multimodal SA. First, the two types of input data are user embedded data and candidate item embedded data. They represent the features of users in social media and the features of candidate items associated with them, respectively. Then, the data flow to the attention capsule layer, which focuses on the more important features by adjusting the weights of the input data. Subsequently, the data undergoes a 3×3 convolution operation for extracting spatial features and a nonlinear transformation by the ReLU activation function. Finally, the processed data enter the FCL for further feature learning. The final output weight coefficients are used in SA or recommendation tasks to help the system accurately determine the importance of each candidate item or its match with the user's emotion [17]. The attention capsule utilizes the output data to adjust the attention bias of the rest of the recommender system, while the attention capsule internally forms the final output by performing multiple convolutions, mappings, full connections, and other changes to the input data [18]. Specifically, the output $v_U(A)$ of the attention capsule is calculated according to Equation (9).

$$v_U(A) = f(v_A, e_1, e_2, \ldots, e_m) = \sum_{j=1}^{m} g(e_j, v_A) e_j = \sum_{j=1}^{m} weight_j e_j \quad (9)$$

In Equation (9), v_A represents the embedding vector, which is capable of fusing the embedding representation of the user's historical behavior with the candidate content. e_j is the candidate feature. $g()$ is the feedforward network. $weight_j$ represents the activation weights, which are calculated by the feedforward network. The input to the attention capsule is the embedding vector of historical behavior and candidate products. The outer product of the two is chosen here, and a convolution kernel is used to convolve these three after splicing and combining them [19]. The output is then connected to a FCL to obtain the magnitude of the weight values. The attention capsule model used in this study sheds the Softmax layer in the traditional attention framework. This means that the obtained sum of weights is not equal to 1. By discarding the Softmax layer, the activation level of the item can be maintained. In other words, a higher weight total indicates a stronger correlation between the item and past behavior, which improves the model's capacity to reflect interest [20]. The study names the proposed model as MSAM.

3 Performance evaluation and experimentation of MSAM

The experiment uses NVIDIA GeForce RTX 3090 GPUs for accelerated computation in hardware with Intel Core i9-10900K processor with 64GB RAM to guarantee the efficient operation of model training. Coupled translation fusion network (CTFN), multimodal InfoMax (MMIM), sentimental words aware fusion network (SWAFN), and MSAM proposed by the study are selected for comparison. The CMU Multimodal Opinion Sentiment and Intensity (CMU-MOSI) and Body-Expression-Audio-Text Dataset (BEAT) are used for the dataset. The Adam optimizer is employed during training, with a starting learning rate of 0.001. The batch size is 32, and the number of training rounds (epoch) is set at 90. The early stopping method is also employed to avoid overfitting. A comparison of Box_Loss convergence curves for different models on different datasets is shown in Figure 6.



(a) CMU-MOSI

(b) BEAT

Figure 6 Comparison of Box_Loss convergence curves for different models on different datasets

Received: July 20, 2025

The variation of Box_Loss with the number of training rounds (Epoch) for different methods on the CMU-MOSI (a) and BEAT (b) datasets is illustrated in Figure 6. On the CMU-MOSI dataset, the proposed MSAM method of the study shows a rapid decrease in Box_Loss with increasing Epoch and eventually reaches a lower level. Compared with the CTFN, SWAFN, and MMIM methods, it shows better loss decreasing trend and lower final value, demonstrating stronger optimization ability. On the BEAT dataset, MSAM also performs well. Box_Loss decreases rapidly and the final value is significantly lower than the other compared methods, further proving that MSAM can effectively reduce the loss on different multimodal SA datasets. It illustrates the method's superiority and efficacy in the multimodal SA task. The study incorporates the new adaptive fusion network (AFN) model to further develop the performance comparison of different SA models as shown in Table 1.

Table 1 Performance comparison of different SA models

Model name	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC	Training fit (%)	Training time (hours)
CTFN	88.723	89.105	87.401	88.228	0.914	0.870	12.75
MMIM	89.341	89.470	88.800	89.134	0.920	0.883	11.68
SWAFN	89.527	89.811	88.950	89.379	0.928	0.888	10.82
AFN	89.833	90.060	89.270	89.664	0.930	0.891	9.98
MSAM	90.285	90.442	89.517	89.980	0.935	0.892	9.56

Table 1 demonstrates the performance comparison of different SA models, including CTFN, MMIM, SWAFN, AFN, and MSAM models. Overall, the models perform differently in each index. The MSAM model performs best in accuracy, precision, recall, F1-score, AUC, and training fit with 90.285%, 90.442%, 89.517%, 89.980%, 0.935, and 0.892, respectively. The training time is also shorter at 9.56 hours. The CTFN model has relatively low indicators, with an accuracy of 88.723%. The training time is the longest, 12.75 hours. The MMIM, SWAFN, and AFN models are in the middle level of each index, and the training time also decreases in order. Figure 7 presents a comparison of various models with respect to efficiency and loss.

Received: July 20, 2025

Volume 38 No. 1s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

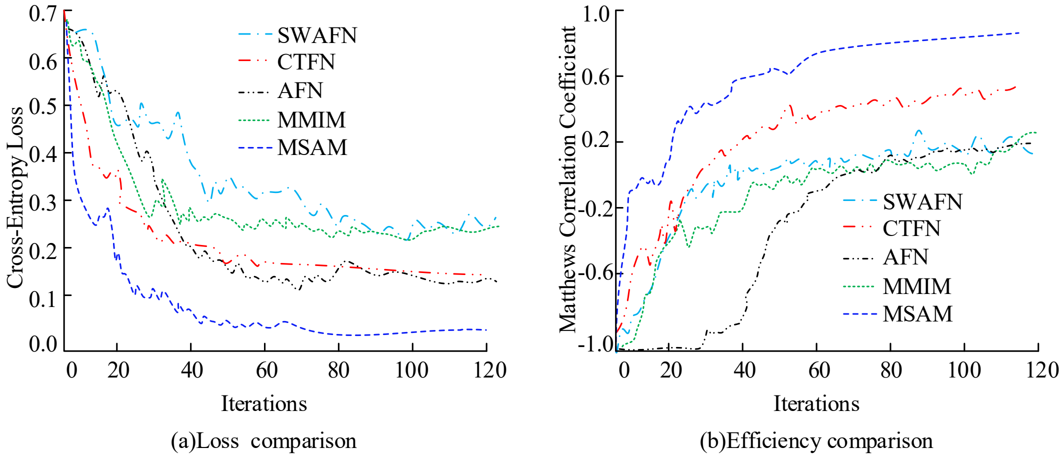


Figure 7 Comparison of different models in terms of loss and efficiency

Figure 7 presents a comparison of various models with respect to efficiency and loss. Figure 7(a) shows the loss comparison. The cross-entropy loss of MSAM decreases rapidly as the iteration increases. Compared with models such as SWAFN, CTFN, AFN, and MMIM, the decreasing trend is more significant and eventually converges to a lower loss value. It shows that MSAM has better convergence in the training process and can reduce the loss more effectively. Figure 7(b) shows the efficiency comparison, the Matthews correlation coefficient of MSAM improves rapidly with the increase in the number of iterations. Moreover, it eventually reaches a higher value, which is significantly better than other models. This suggests that MSAM is superior in terms of model performance and efficiency since it can more effectively capture data features and enhance the model's overall performance. Table 2 displays the MSAM thorough performance evaluation.

Table 2 Comprehensive performance evaluation of MSAM

Model name	Inference speed (ms)	Memory usage (GB)	Training stability (%)	Model scalability (%)	Prediction accuracy (%)	Error rate (%)	Training time (hours)
CTFN	75.214	3.283	91.472	88.125	86.302	13.187	12.85
MMIM	68.359	2.762	90.415	89.420	87.647	11.438	11.92
SWAFN	82.371	3.054	92.031	90.526	89.024	10.302	10.65
AFN	70.725	2.958	93.065	91.142	90.561	9.473	10.23
MSAM	64.887	2.487	94.209	92.476	91.748	8.196	9.37

Table 2 demonstrates the comprehensive performance evaluation of the modal polymorphic SA model. The MSAM model performs best on all metrics. It has the fastest inference speed (64.887ms) and the least memory usage (2.487GB). It has the highest

Received: July 20, 2025

training stability (94.209%), scalability (92.476%), and prediction accuracy (91.748%). Moreover, it has the lowest error rate (8.196%) and training time (9.37 hours). The other models differ in each metric.The CTFN model is relatively behind in terms of inference speed, memory usage, and training stability, as well as lower metrics such as accuracy rate and error rate. Table 3 displays the multidimensional performance evaluation of the SA model in the emotion monitoring task.

Table 3 Multidimensional performance evaluation of SA model in emotion monitoring task

Metric name	MSAM	CTFN	AFN	SWAFN	MMIM	MSAM
Emotion analysis response time (ms)	241.352	265.987	235.671	278.941	259.384	248.329
Data preprocessing time (ms)	120.236	132.146	118.589	140.882	125.493	118.997
Model inference time (ms)	80.557	91.346	75.469	96.728	88.133	84.129
Sentiment classification accuracy (%)	91.853	89.442	92.431	88.512	90.128	91.208
Positive sentiment proportion (%)	54.602	51.189	55.834	49.723	52.013	53.274
Negative sentiment proportion (%)	38.265	40.138	37.546	42.396	39.228	38.715
Neutral sentiment proportion (%)	7.133	8.673	6.620	7.881	8.759	8.011
Emotion analysis error rate (%)	8.146	9.135	7.568	10.824	8.745	8.339
Real-time monitoring system stability (%)	94.287	91.146	93.142	90.845	92.320	94.127
Emotion feedback adjustment time (ms)	312.741	330.493	310.568	355.733	338.902	321.345
Training data fit (%)	98.246	97.345	98.653	96.529	97.981	98.128
System resource utilization rate (%)	82.134	87.536	80.439	88.253	83.961	84.728
Emotion analysis output time (ms)	50.183	58.361	45.896	62.485	55.723	52.491

Table 3 demonstrates the multidimensional performance evaluation of MSAM and other SA models in the emotion monitoring task. The MSAM model performs better in sentiment classification accuracy (91.853%), real-time monitoring system stability (94.287%), and training data fit (98.246%). However, it is not all leading in terms of emotion analysis response time (241.352ms) and data preprocessing time (120.236ms). Overall, the

Received: July 20, 2025

performance of the models varies in different dimensions, and the MSAM model has some advantages in most of the key indicators. However, there is still room for improvement in some of the timeliness indicators.

4.Conclusion

With the explosive growth of multi-platform UGC, multimodal SA is becoming more and more important in understanding complex emotional expressions. However, traditional methods face challenges such as insufficient feature fusion and insufficient focusing of AMs. To this end, the research constructed multi-source heterogeneous data collection architecture, fused text, image, and other multimodal data, and proposed MSAM model. Text features were extracted by BERT, and sequence dependency was modeled by combining LSTM with two-layer Transformers module. Moreover, the sentiment loss function was introduced to guide the generation of responses with sentiment features. The experimental results indicated that MSAM significantly outperformed the comparison models on CMU-MOSI and BEAT datasets. The accuracy reached 90.285%, the F1-score was 89.980%, the AUC reached 0.935, and the training time was only 9.56 hours. In the comprehensive performance evaluation, its inference speed was 64.887ms, memory usage was 2.487GB, prediction accuracy was 91.748%, real-time monitoring system stability was 94.287%, and training data fit was 98.246%. It demonstrated efficient feature extraction and sentiment classification. However, the model still had room for improvement in terms of emotion analysis response time (241.352ms) and deep fusion of multimodal data (e.g., video, audio). Future research aims to optimize the model structure to enhance real-time performance and expand to cross-modal feature interaction modeling. To increase the model's flexibility in challenging emotional contexts, it will also investigate landing applications in social opinion tracking, intelligent customer service, and other scenarios.

References

[1] Choudhary M, Chouhan S S, Rathore S S. Beyond Text: Multimodal Credibility Assessment Approaches for Online User-Generated Content[J]. ACM Transactions on Intelligent Systems and Technology, 2025, 15(5): 1-33.

[2] Omar K, Sakr R H, Alrahmawy M F. An ensemble of CNNs with self-attention mechanism for DeepFake video detection[J]. Neural Computing & Applications, 2024, 36(6): 2749-2765.

[3] Bansal S, Kumar M, Raghaw C S, Kumar N. Sentiment and hashtag-aware attentive deep neural network for multimodal post popularity prediction[J]. Neural Computing and Applications, 2025, 37(4): 2799-2824.

[4] Das R, Singh T D. Multimodal sentiment analysis: A survey of methods, trends, and

Received: July 20, 2025

challenges[J]. ACM Computing Surveys, 2023, 55(13s): 1-38.

[5] Singh U, Abhishek K, Azad H K. A survey of cutting-edge multimodal sentiment analysis[J]. ACM Computing Surveys, 2024, 56(9): 1-38.

[6] Ghorbanali A, Sohrabi M K. A comprehensive survey on deep learning-based approaches for multimodal sentiment analysis[J]. Artificial Intelligence Review, 2023, 56(Suppl 1): 1479-1512.

[7] Sun L, Lian Z, Liu B, Tao J. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis[J]. IEEE Transactions on Affective Computing, 2023, 15(1): 309-325.

[8] Li Y, Lan X, Chen H, Lu K, Jiang D. Multimodal pear chain-of-thought reasoning for multimodal sentiment analysis[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2025, 20(9): 1-23.

[9] Wang P, Zhou Q, Wu Y, Chen T, Hu J. DLF: Disentangled-Language-Focused Multimodal Sentiment Analysis[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(20): 21180-21188.

[10] Wu J, Zhu T, Zhu J, Li T, Wang C. A optimized bert for multimodal sentiment analysis[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 19(2s): 1-12.

[11] Gou J, Sun L, Yu B, Wan S, Tao D. Hierarchical multi-attention transfer for knowledge distillation[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 20(2): 1-20.

[12] Kumar S, Solanki A. An abstractive text summarization technique using transformer model with self-attention mechanism[J]. Neural Computing and Applications, 2023, 35(25): 18603-18622.

[13] Xia Y, Ding D, Chang Z, Li F. Joint deep networks based multi-source feature learning for QoS prediction[J]. IEEE Transactions on Services Computing, 2022, 15(4): 2314-2327.

[14] Wang M, Li X M, Zhang X, Zhang Y. Hierarchical graph attention network with pseudo-metapath for skeleton-based action recognition[J]. Neurocomputing, 2022, 501(8): 822-833.

[15] Xiao G, Wei Y, Yao H, Deng W, Xu J, Pan D. Hierarchical broad learning system for hyperspectral image classification[J]. IET Image Processing, 2022, 16(2): 554-566.

[16] Sun C, Li C, Lin X, Zheng T, Meng F, Rui X, Wang Z. Attention-based graph neural networks: A survey[J]. Artificial Intelligence Review, 2023, 56(Suppl 2): 2263-2310.

[17] Meng K, Dong X, Shan H, Xia S. Multiscale hierarchical attention fusion network for edge detection[J]. International Journal of Ad Hoc and Ubiquitous Computing, 2023,

42(1): 1-11.

[18] Martin R J, Oak R, Soni M, Mahalakshmi V, Soomar A M, Joshi A. Fusion-based Representation Learning Model for Multimode User-generated Social Network Content[J]. ACM Journal of Data and Information Quality, 2023, 15(3): 1-21.

[19] Lei Y, Qu K, Zhao Y, Han Q, Wang X. Multimodal sentiment analysis based on composite hierarchical fusion[J]. The Computer Journal, 2024, 67(6): 2230-2245.

[20] Vinitha V, Bargavi S M. Transforming recommender system by integrating attention-based neural network for multimodal sentiment analysis using artificial intelligence[J]. International Journal of Computer Applications in Technology, 2024, 74(1-2): 136-145.

Received: July 20, 2025