Journal of Nonlinear and Convex Analysis
Volume 26, Number 6, 2025, 1785–1796

Yokohama Publishers
ISSN 1880-5221 ONLINE JOURNAL

SENTIMENT ANALYSIS BASED ON MULTIMODAL SENTIMENT MULTIHEAD ATTENTION FUSION

BOXIONG CHEN, CAIMAO LI*, BIYUAN YAO, AND SHAOFAN CHEN

Abstract. Sentiment analysis aids in understanding emotions and social interactions, with applications in areas like public opinion monitoring and emotionally intelligent interactions. However, most research focuses on single-modal data, missing the richness of multimodal information. In addition, single-head attention mechanisms struggle with multimodal data integration. To address this, we propose using multi-head attention for feature extraction and dimensionality reduction, retaining key sentiment information. First, feature extraction with the multimodal multi-head attention mechanism is performed for projection into the vector subspace to decrease the dimensionality of the features and retain the most dominant sentiment feature information for subsequent analysis. Second, the multimodal features are computed using the labels of the emotion dataset and projected into the vector space to achieve the fusion of multimodal features, thus effectively integrating the information of different perceptual channels from text, audio and video. Finally, experiments on the multimodal datasets of the Multimodal Corpus of Sentiment Intensity (CMU-MOSI) dataset and the Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset are carried out to verify the effectiveness and performance of the multimodal attention mechanism in the multimodal sentiment analysis task. The experimental results show that the multimodal sentiment analysis approach incorporating the multi-head attention mechanism improves accuracy, sentiment score, precision, and recall metrics by at least 1%. With an execution time of around 60 minutes, it is the most efficient among all models.

1. Introduction

Sentiment analysis is a key task in Natural Language Processing (NLP), but as datasets grow, single-modal methods are becoming inadequate. Some researchers have explored opinion mining and user behavior analysis [1], but most focus on text, requiring multiple models for strong results. Later studies used multimodal co-training to analyze sarcasm, achieving better sentiment analysis outcomes [22]. The increasing dataset size highlights the limitations of single-sentiment analysis.

In multimodal sentiment analysis, large datasets require models to be efficient and stable, with intermediate mechanisms seamlessly integrated. Researchers proposed a unified framework for analyzing missing and unaligned patterns [6]. However, without attention mechanisms, fusion analysis results were unstable. Transformer-based interaction modules were later introduced but are limited by input sequence

2020 *Mathematics Subject Classification.* 68T50, 68T07.

Key words and phrases. Multi-situational, sentiment analysis, multihead attention mechanisms.

This work was supported by the Hainan Provincial Natural Science Foundation of China (Grant No. 625QN269).

*Corresponding author.

length and lack contextual accuracy [3]. Subsequent studies combined coding techniques and multiple attention mechanisms for sequence extraction and sentiment analysis [21]. While tag coding helps manage missing data, it can result in information loss or inaccuracy. Multi-head attention with a fusion graph grid improved accuracy but increased complexity and computational cost [17]. This limits model scalability and real-time performance, especially with larger datasets.

To address these issues, we propose a lightweight multimodal sentiment analysis model, namely the multimodal multi-head attention mechanism perceptual fusion model. It uses multi-head attention to extract key features from text, audio, and video. The model randomly selects two modes, then integrates a third mode, combining multimodal fusion with attention mechanisms to preserve emotional features. Finally, the features of all three modes are fused for sentiment analysis. The prime contributions of the paper are as follows:

- Multimodal feature extraction is carried out using multi-head attention mechanism and mapped to the vector quantum space to cut down the feature dimension while retaining the main emotional feature information.
- The fusion of multimodal multi-head attention multi-channel data is realized by calculating the emotion labels of the data and fusing different channels from text, audio, and video with different modalities for analysis.
- After experiments on the multimodal datasets of CMU-MOSI and CMU-MOSEI, compared with the previous baseline model, our proposed model has achieved significant improvement in sentiment analysis tasks, which well proves the effectiveness of the method.

The rest part is structured as follows. Section 2 reviews related work, focusing on feature extraction methods and fusion strategies to enhance model performance. Section 3 details the multimodal feature fusion using the multi-head attention model, including the design of attention mechanisms and the integration of multi-head attention into feature extraction and fusion. Section 4 validates method using CMU-MOSI and CMU-MOSEI datasets. Section 5 concludes paper.

2. Related work

2.1. **Text feature analysis.** In text modality research, attention mechanism and Long Short-Term Memory (LSTM) network[7] are combined to link context and identify emotion-related words. This proved the attention mechanism's importance in word-context relationships. For text feature extraction and analysis, the Bidirectional Encoder Representations from Transformers (BERT) [8] model is adopted to extract rich text features, playing a key role in multimodal sentiment analysis.

2.2. **Audio feature analysis.** Audio feature analysis involves extracting features from audio signals. Some researchers suggest using end-to-end deep learning with neural networks to directly learn emotional representations from raw audio[13], bypassing manual feature design and enhancing emotion recognition. This approach offers more accurate features for multimodal sentiment analysis. We use the Collaborative Voice Analysis Repository for Speech Technologies (COVAREP)[10] for audio feature extraction to support multimodal sentiment analysis research.

2.3. **Video Feature Analysis.** Video feature analysis involves extracting and understanding features from video data to support various analytics and sentiment analysis tasks. Some researchers use images to align with text in social media emotion analysis, highlighting key sentences [11]. This study shows that images can serve a similar role as text. For video feature extraction, we use facet[14], which provides detailed feature representations of text and images in video, enhancing our understanding and improving accuracy in multimodal sentiment analysis.

2.4. **Multimodal sentiment analysis.** With the rise of diverse datasets, single-modal sentiment analysis no longer suffices for multimodal applications, making multimodal sentiment analysis more complex. This approach generally involves two main components: feature learning and modal fusion. Feature learning typically involves supervised methods on labeled datasets to learn useful representations for tasks like classification or regression. Common methods include Convolutional Neural Networks (CNN) [2] and Recurrent Neural Networks (RNN) [9]. However, feature-based approaches can be time-consuming and labor-intensive, with potential issues of feature bias. Modal fusion is key in multimodal sentiment analysis. Researchers use models to handle modal uncertainty and specific feature representations for better integration of different modalities [5]. Recent approaches, like the two-by-two cross-modal insertion attention mechanism, enhance interaction between multimodal sequences [12]. Despite this, existing methods have limitations. Using multi-head attention to dynamically weight modes based on relevance can further enhance multimodal sentiment analysis performance [15].

3. Muiltimodal Sentiment Multi-head Attention Fusion(MSMAF)

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{3.1}$$

3.1. **Modal feature extraction.** Features can be extracted by calculating the statistical characteristics of the modal data (such as mean, variance, maximum, minimum, etc.), specifically, this formula counts the total sum of the data points x_i in the data sample and divides it by the total number of data samples entered N. And that gives you the mean of this set of data μ.

3.2. **Fusion.** Following feature extraction from the three modalities, we proceed to multimodal fusion, utilizing a multi-head attention mechanism. As shown in Figure 1, extracting features from text, audio, and video fully leverages multimodal correlations, enhancing emotion recognition performance. Feature vectors from these modalities are first aligned in dimension via a linear layer, then fused by integrating two modalities and adding a third. The formula for this process is as follows:

$$A \oplus B\left(\left[A : B'\right] + Segment\right), \tag{3.2}$$

$$B' = Linear\left(B\right) \tag{3.3}$$

where A⊕B is defined as "merging A and B". [A : B′] represents the joining of the transpose of matrix A and matrix B to form a new matrix, obtained by Linear projection B in equation (3.3).

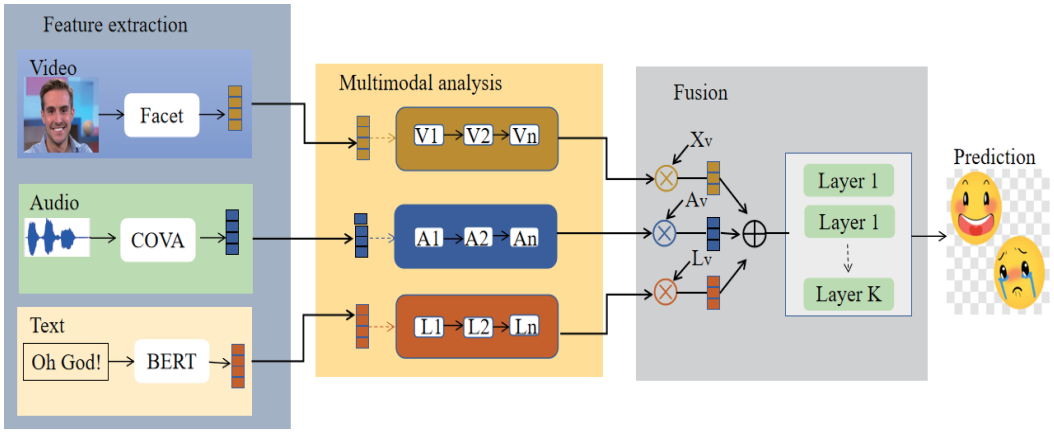

FIGURE 1. Our proposed of MSMAF

After combining A and B′ based on the sequence length, a segment is introduced to enhance the accuracy of the analysis. The approximate fusion model is shown in Figure 1.

3.3. **Multihead attention.** The multi-head attention mechanism boosts the model's ability to understand and integrate relationships across various modalities.

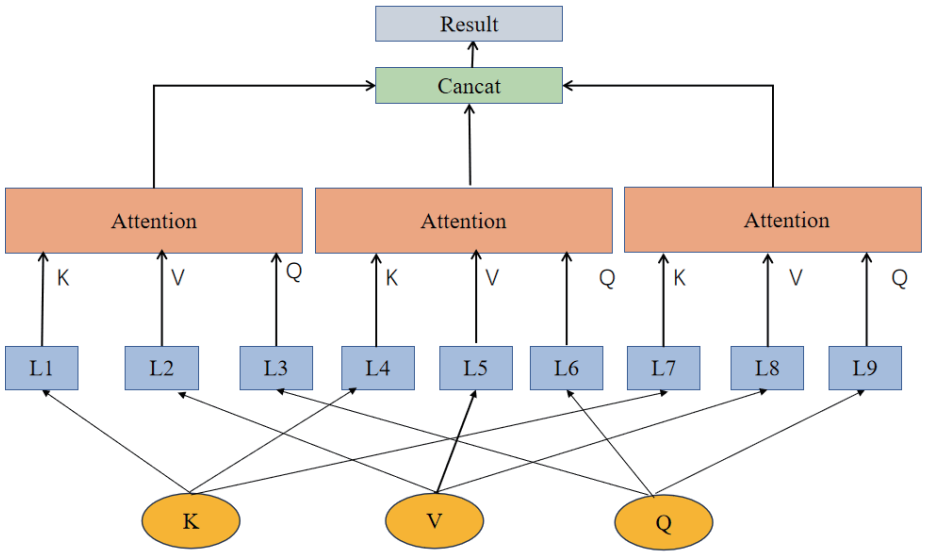

FIGURE 2. The principle of multi-head attention mechanism

In Figure 2, the vectors Q, K, and V represent the query, key, and value vectors, respectively, with their attention weights calculated. The outputs of each attention

head are combined to capture diverse attention results. The multi-head attention mechanism, an advanced version of self-attention, uses several sets of weight matrices to learn various contextual influences. After training, these matrices project into different subspaces, allowing each attention head to focus differently and enhancing the model's capacity.

$$MuiltiHead(Q,K,V) = concat\left(head_1, head_2, \ldots, head_n\right) W^o \tag{3.4}$$

where n is the number of attention $head_n$ and W^o is the weighted matrix from which the matrix of weights for calculating a particular head can be derived:

$$head_i = softmax\left(\frac{QW_i^Q(KW_i^K)^T}{\sqrt{d_k}}\right)V. \tag{3.5}$$

Q, K, V denote the linear transformations of Query, Key and Value respectively. W_i^Q, W_i^K are the linear transformation matrices of each attention $head_i$, and d_k is the dimensions of the query or key.

3.4. **MSMAF.** By introducing the multi-head attention mechanism, we propose the MSMAF model. In the process of each interaction between receiving data and attention, MSMAF can receive data information from each different mode and dynamically integrate emotional labels and key information to be weighted from different modes by using the multi-head attention mechanism. Including text, audio, and video, the formula is shown below:

$$C = MultiHead\left(\frac{XW_QW_i^TK^T}{\sqrt{d}} + U(MSMAF)XW_V\right). \tag{3.6}$$

In Eq.(3.6), C represents the result of prediction, i represents how many parameters need to be included in the formula for calculation, d represents the dimensionality proposed by the previous features, and U represents the conversion of MSMAF into a weight matrix, which facilitates the effective calculation of the formula.

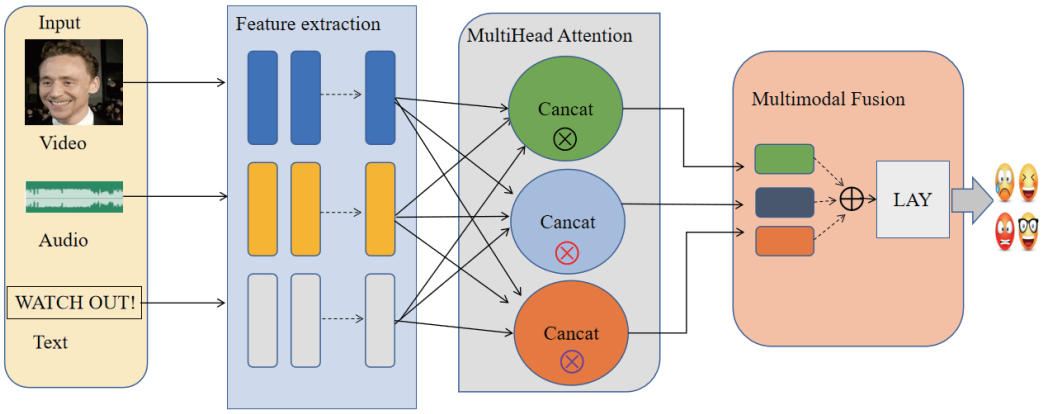

FIGURE 3. A framework for multimodal fusion multi-head attention mechanism

In Figure 3, with the application of multi-head attention mechanism, all attention heads will focus on different key modal information, enabling the model to better understand the overall multimodal representation.

4. Experimental design

4.1. **Experimental Environment.** Hardware: Intel Xeon 4210R CPU, Nvidia RTX 3060 GPU, 16 GB RAM. Software: Linux 64-bit OS, PyCharm, and PyTorch for deep learning.

4.2. **Dataset.** In this paper, we perform experiments using two multimodal sentiment analysis datasets: the CMU-MOSI [20] dataset and the CMU-MOSEI [19] dataset, which correspond to experimental datasets that allow for more direct and efficient experimental evaluation. The CMU-MOSI dataset includes video and blog content from YouTube, featuring multimodal data–video, audio, and text–on various emotions and themes. The CMU-MOSEI dataset, with over 1,000 video clips, is larger and covers a broader range of topics than CMU-MOSI, making it more diverse and representative of real-world emotional expression.

Table 1. Introduction to experimental datasets

Dataset	Train	Valid	Test	Total
CMU-MOSI	129	220	646	2162
CMU-MOSEI	17216	1621	4792	23629

In Table 1, "Train" indicates the experimental training set, "Valid" indicates the validation set, "Test" indicates the test set, and "Total" indicates the overall sample size of the dataset.

4.3. **Experimental evaluation indicators.** The model's performance was evaluated using accuracy and the F1 score. Accuracy measures how often the model correctly predicts sentiment in data fragments, while the F1 score, the harmonic mean of precision and recall, assesses the model's overall sentiment analysis performance. Accuracy indicates the proportion of true positives among predicted positives, and recall measures the percentage of actual positives correctly identified by the model. Execution time was used to compare the model's efficiency.

4.4. **Feature extraction.** The emotion-related features of the three models are extracted, to support the subsequent multimodal emotion fusion analysis.

Table 2. The three modalities dimensions

Dataset	L	A	V
CMU-MOSI	768	74	34
CMU-MOSEI	768	74	35

The experiment extracts features from text (L), audio (A), and video (V). Text features use a pre-trained BERT model, audio features are obtained via COVAREP, and video features come from Facet. These features are labeled L, A, and V, respectively, as shown in Table 2.

4.5. **Baseline model.** We compared the MSMAF model to other deep learning baselines in sentiment analysis, utilizing all three modalities: text, audio, and video, to validate its effectiveness.

- Multi-grained Attention Network (MGAN) [4]: The interaction between sentence and lexical context is learned through coarse-grained and fine-grained attention mechanisms.
- Modality-invariant and Specific Representations for Multimodal Sentiment Analysis (MISA) [5]: Information from the three modes is decomposed into modal invariance and specificity to identify commonalities and characteristics, reducing the gap between modes.
- Multi-Interactive Memory Network (MIMN) [16]: Learning the interaction of information between specific text and image modalities using a dual memory network.
- Visual Aspect Attention Network (VistaNet) [11]: Image-text fusion of emotion classification and three-layer pattern architecture, image pattern as the alignment vector to emphasize the important information of the sentence.
- Tensorfusion Network (TFN) [18]: This model learns to dynamically interact with information across the three modes, aggregating interactions from endpoint to endpoint.

4.6. **Analysis of experimental results.** Through experimental comparisons with other baseline models, all experimental results show that MSMAF outperforms the baseline model and improves accuracy, F1 emotion scores and other indicators through its attention mechanism. It is also superior in speed and improves the efficiency of analysis on multimodal sentiment datasets.

Table 3. Comparison of CMU-MOSI multimodal results

	Accuracy (%)	F1 (%)	Precision (%)	Recall (%)	Time (min)
MGAN	71.52	71.32	71.45	71.12	62.35
MIMN	73.02	73.2	72.95	72.88	63.41
VistaNet	75.91	75.85	75.77	75.65	67.61
TFN	78.81	78.74	78.80	78.71	72.41
MISA	81.12	81.35	81.09	81.31	60.24
MSMAF	**82.94**	**82.89**	**82.79**	**83.27**	**53.58**

Table 3 shows the comparison between MSMAF and other baseline models. Experiments are conducted based on CMU-MOSI multimodal sentiment analysis dataset, including accuracy, F1, precision, recall, and time, where black and bold words represent better experimental results. After experimental comparison between MSMAF and other baseline models, it can be intuitively known that the proposed model is superior to other baseline models in experimental index. Among the baseline models, MSMAF outperforms others by improving accuracy, F1 emotion scores, and other metrics by over 1 percentage points with the attention mechanism. Additionally, it is the fastest, demonstrating improved sentiment analysis efficiency.

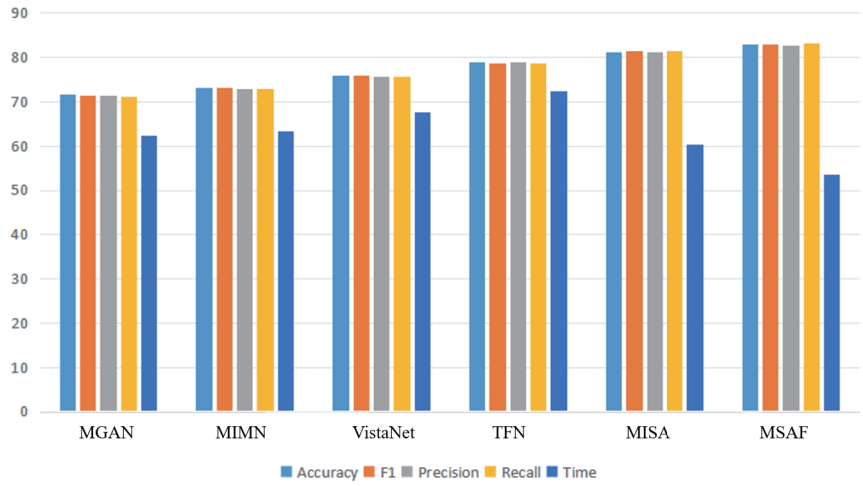

FIGURE 4. CMU-MOSI experimental data graph

As shown in Figure 4, the experiment uses the bar chart to visually show that MSFA improves the accuracy, F1 emotion scores and other indicators by at least 1 percentage point through the fusion of multi-head attention mechanism through the multimodal sentiment analysis model, and the experimental results are significantly better than the performance of other models in this experiment. In addition, it is the fastest sentiment analysis model and has proven that it can improve the efficiency of sentiment analysis on complex multimodal sentiment datasets.

TABLE 4. Comparison of the CMU-MOSEI multimodal results

	Accuracy (%)	F1 (%)	Precision (%)	Recall (%)	Time (min)
MGAN	75.61	75.51	75.69	75.49	76.54
MIMN	77.51	77.38	77.43	77.24	74.61
VistaNet	78.34	78.24	78.41	78.31	81.45
TFN	80.23	80.13	80.11	80.14	79.62
MISA	81.61	81.54	81.64	81.51	72.48
MSMAF	**83.75**	**83.65**	**83.43**	**83.44**	**64.59**

Table 4 shows a comparison between MSMAF and other baseline models. The experiment is based on the larger and more complex CMU-MOSEI multimodal sentiment analysis dataset, including accuracy, F1, precision, recall and time, where black and bold represent better experimental results. Through the experimental comparison between MSMAF and other baseline models, it can be intuitively known that the model proposed in this paper is superior to other baseline models in experimental indexes. In the baseline model, MSMAF outperformed other models by improving accuracy, F1 emotion scores, and other metrics by nearly 2 percentage by using attention mechanisms. In addition, the running efficiency of the proposed model is the fastest. Through the comparison of running time, it is proved that the efficiency of sentiment analysis is improved.

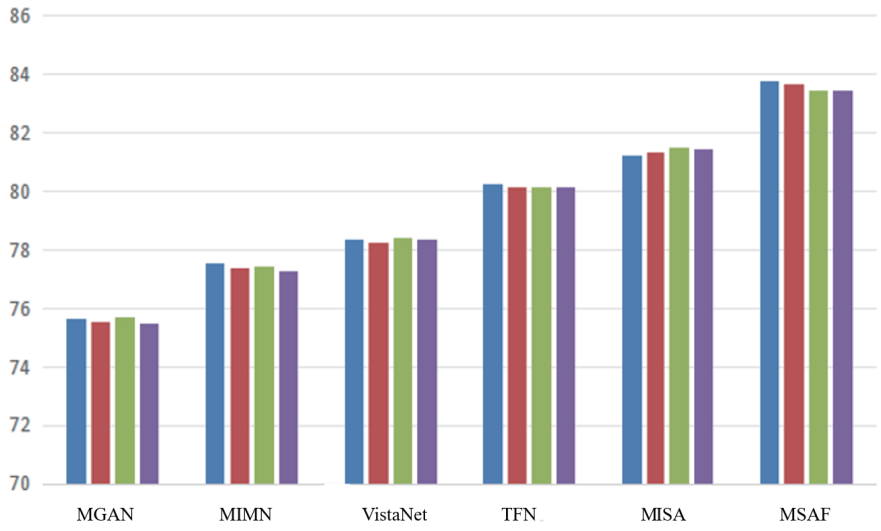

FIGURE 5. CMU-MOSEI experimental data graph

It can be seen that after adding the attention mechanism to MSMAF, the accuracy, F1 emotion scores and other indicators have increased by nearly 2 percentage points compared to other baseline models. Moreover, among the various models, the model studied in this paper takes the least time and improves the efficiency of emotion analysis. The visualized data is presented in Figure 5.

4.7. **Ablation experiment.** Since the data of CMU-MOSEI is more abundant, we plan to conduct ablation experiments on this dataset, compare the complete MSMAF with the model after extracting each mode and the missing key multi-head attention mechanism in various aspects, and observe whether there are changes in the experimental indicators.

TABLE 5. Ablation experiments on the CMU-MOSEI dataset

	Accuracy (%)	F1 (%)	Precision (%)	Recall (%)	Time (min)
MSMAF-L	65.32	64.45	64.89	64.34	54.68
MSMAF-A	61.96	61.32	62.13	62.09	51.25
MSMAF-V	68.41	67.59	68.21	68.04	**41.54**
MSMAF-ATT	72.22	71.69	72.15	71.39	58.78
MSMAF	**83.75**	**83.65**	**83.43**	**83.44**	64.59

In terms of the ablation experiment module, we identified four missing modules for comparison, namely, MSMAF-L containing only the text part of the analysis data, MSMAF-A containing only the audio part of the analysis data, MSMAF-V containing only the video part of the analysis data, and the fusion model MSMAF-ATT after removing the multiple attention mechanism. The four missing modules are compared with the complete MSMAF, and the results are shown in Table 5.

As can be seen from the above table, when one of the three modules is indeed missing, the experimental data of the model proposed in this paper on the prediction of sentiment analysis plummets compared to the previous one. The reason is that the multimodal sentiment analysis model is not able to play its corresponding advantages when a certain module is missing, and also corresponds to the lack of certain key sentiment information when a certain module is missing, which leads to the experimental index of sentiment analysis is not high.

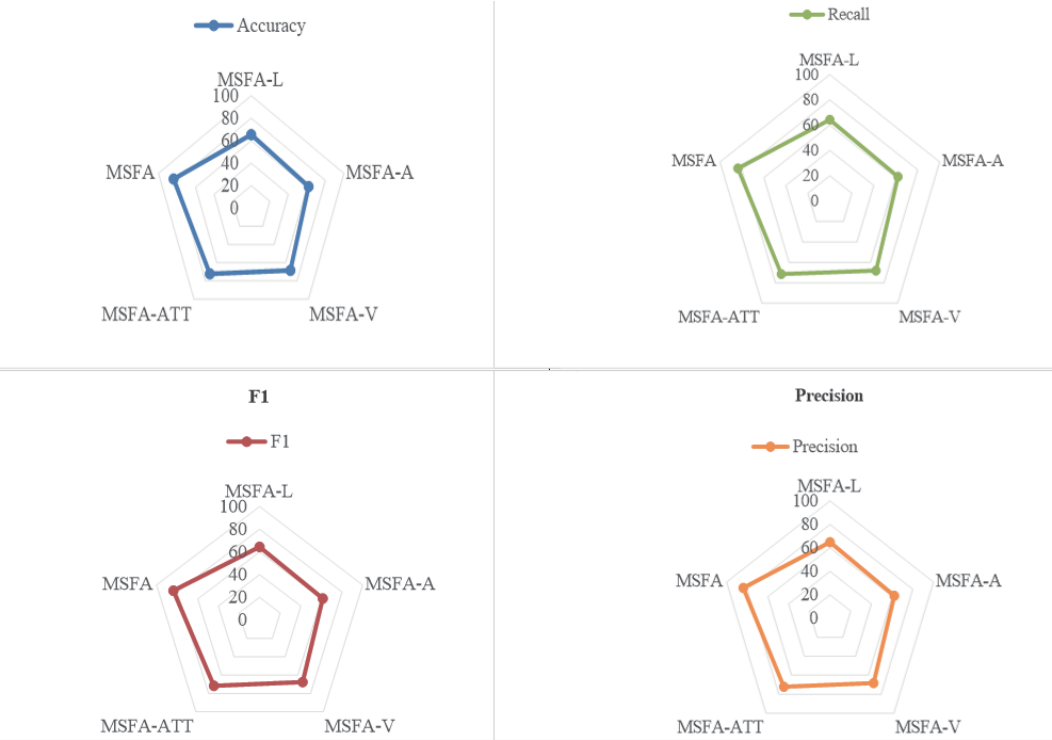

FIGURE 6. Comparison of ablation experiment data

In Figure 6, by removing some modules in the whole experiment one by one, such as retaining only the text module, removing multi-head attention mechanism, etc., the MSMAF model proposed in this paper has the most stable performance. In the ablation experiment, if only one module of the three modules is used, the experimental effect is suboptimal, and the average efficiency is only about 60%. When the three modules are present at the same time but the multi-head attention mechanism is missing, the average efficiency reaches 70The above categories of ablation experiments summarize the experimental performance of our proposed model approach on multimodal sentiment analysis datasets. By integrating multi-head attention mechanisms, the relevant performance and indicators of multimodal sentiment analysis have been improved, which indicates that multi-head attention mechanisms can effectively help multimodal models obtain relevant important information between different modes, and help improve the accuracy and effect of sentiment analysis.

5. Summary

We perform sentiment analysis on a multimodal sentiment dataset, utilizing three modalities for feature extraction and incorporating the MSMAF model for multimodal sentiment analysis. The performance of the model is further enhanced by adding a multi-head attention mechanism, which obtains higher accuracy compared to other baseline models and achieves some new enhancements in various experimental metrics. More efficient algorithms will be designed and integrated into the model in subsequent studies to build a richer and more efficient model for multi-sentiment analysis. This attitude of continuous improvement and progress is highly commendable and helps to continuously improve the performance and effectiveness of multimodal sentiment analysis tasks.

References

[1] A. Belhadi, Y. Djenouri, A. N. Belbachir, T. Michalak and G. Srivastava, *Shapley visual transformers for image-to-text generation*, Applied Soft Computing **166** (2024): 112205.

[2] E. D. Cherpanath, P. R. F. Nasreen, K. Pradeep, M. Menon and V. S. Jayanthi, *Food image recognition and calorie prediction using faster R-CNN and mask R-CNN*, in Proceedings of the 9th International Conference on Smart Computing and Communications (ICSCC), IEEE, 2023, pp. 83–89.

[3] G. Degottex, J. Kane, T. Drugman, T. Raitio and S. Scherer, *COVAREP – A collaborative voice analysis repository for speech technologies*, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2014, pp. 960–964.

[4] F. F. Fan, Y. S. Feng and D. Y. Zhao, *Multi-grained attention network for aspect-level sentiment classification*, in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2018, pp. 3433–3442.

[5] D. Hazarika, R. Zimmermann and S. Poria, *Misa: Modality-invariant and-specific representations for multimodal sentiment analysis*, in Proceedings of the 28th ACM International Conference on Multimedia (ACM MM), Association for Computing Machinery, 2020, pp. 1122–1131.

[6] R. He, W. S. Lee, H. T. Ng and D. Dahlmeier, *Exploiting document knowledge for aspect-level sentiment classification*, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, 2018, pp. 579–585.

[7] R. Huan, G. Zhong, P. Chen and R. Liang, *UniMF: A unified multimodal framework for multimodal sentiment analysis in missing modalities and unaligned multimodal sequences*, IEEE Transactions on Multimedia **26** (2024), 5753–5768.

[8] P. P. Liu, X. Zheng, H. Li, J. Liu, Y. Ren, H. Zhu and L. Sun, *Improving the modality representation with multi-view contrastive learning for multimodal sentiment analysis*, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.

[9] S. Renjith and R. Manazhy, *Indian Sign Language Recognition: A comparative analysis using CNN and RNN models*, in Proceedings of the International Conference on Circuit Power and Computing Technologies (ICCPCT), IEEE, 2023, pp. 1573–1576.

[10] M. H. Phan and P. Ogunbona, *Modelling context and syntactical features for aspect-based sentiment analysis*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, 2020, pp. 3211–3220.

[11] Q. T. Truong and H. W. Lauw, *VistaNet: Visual aspect attention networkfor multimodal sentiment analysis*, in Proceedings of the AAAI Conference on Artificial Intelligence, AAAI, 2019, pp. 305–312.

[12] Y. H. H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.P. Morency and R. Salakhutdinov, *Multimodal transformer for unaligned multimodal language sequences*, in Proceedings of the 57th Annual

Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, 2019, pp. 6558–6569.
[13] P. Tzirakis, J. Zhang and B. W. Schuller, *End-to-End speech emotion recognition using deep neural networks*, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 5089–5093.
[14] D. Vandic, S. Aanen, F. Frasincar and U. Kaymak, *Dynamic facet ordering for faceted product search engines*, IEEE Transactions on Knowledge and Data Engineering **29** (2017), 1004–1016.
[15] Y. Wang, Y. Xie, X. Ji, Z. Liu and X. Liu, *RacPixGAN: An enhanced sketch-to-face synthesis GAN based on residual modules, multi-head self-attention mechanisms, and CLIP loss*, in Proceedings of the 4th International Conference on Electronic Communication and Artificial Intelligence (ICECAI), IEEE, 2023, pp. 336–342.
[16] N. Xu, W. J. Mao and G.D. Chen, *Multi-interactive memory network for aspect based multimodal sentiment analysis*, in Proceedings of the AAAI Conference on Artificial Intelligence, AAAI, 2019, pp. 371–378.
[17] X. Xue, C. Zhang, Z. Niu and X. Wu, *Multi-level attention map network for multimodal sentiment analysis*, IEEE Transactions on Knowledge and Data Engineering **35** (2023), 5105–5118.
[18] A. Zadeh, M. H. Chen, S. Poria, E. Cambria and L. P. Morency, *Tensorfusion network for multimodal sentiment analysis*, in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2017, pp. 1114–1125.
[19] A. Zadeh, P. P. Liang, S. Poria, E. Cambria and L. P. Morency, *Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph*, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, 2018, pp. 2236–2246.
[20] A. Zadeh, R. Zellers, E. Pincus and L. P. Morency, *MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos*, IEEE Intelligent Systems **31** (2016), 82–88.
[21] J. Zeng, J. Zhou and T. Liu, *Robust Multimodal Sentiment analysis via tag encoding of uncertain missing modalities*, IEEE Transactions on Multimedia **25** (2023), 6301-6314.
[22] G. Zhu, E. Deng, Z. Qin, F. Khan, W. Wei, G. Srivastava, H. Xiong and S. Kumari, *Cross-modal interaction and multi-source visual fusion for video generation in fetal cardiac screening*, Information Fusion **111** (2024): 102510.

Manuscript received April 20, 2024
revised September 23, 2024

B. Chen
Hainan University, Haikou 570228, China
E-mail address: 15799048390@163.com

C. Li
Hainan University, Haikou 570228, China
E-mail address: lcaim@hainanu.edu.cn

B. Yao
Hainan University, Haikou 570228, China
E-mail address: byyao@hainanu.edu.cn

S. Chen
Hainan University, Haikou 570228, China
E-mail address: 990588@hainanu.edu.cn