Proceedings of CONF-CDS 2025 Symposium: Data Visualization Methods for Evaluation
DOI: 10.54254/2755-2721/2025.PO24936

Multimodal Sentiment Analysis Using Pre-trained Models: A Study of Few-Shot learning and Social Media Scenarios

Fazong Lin[1*], Junhao Yu[2], Ziqi Ye[3]

[1]*Faculty of Data Science, City University of Macau, Macau, China*
[2]*School of International Education, Jiangsu University of Technology, Changzhou, China*
[3]*Faculty of Science and Technology, Beijing Normal-Hong Kong Baptist University, Zhuhai, China*
**Corresponding Author. Email: D22090104054@cityu.edu.mo*

Abstract. Facing the proliferation of social media and user-published content, traditional single sentiment analysis cannot capture users' emotional expressions well, and multi-modal sentiment analysis (MSA) can solve this problem well. Multi-modal sentiment analysis improves the system's ability to understand complex emotional data through the integration of multi-modal tasks such as audio, image, and video. This paper systematically analyzes five representative multi-modal sentiment analysis pre-training models in recent years, focusing on the model structure, integration methods, adaptability and performance of these five models in small sample-oriented and social media scenarios. Through the analysis and summary of MGT, SensoryT5, AMP, MMEN and VLP2MSA models, these models have shown good robustness and generalization in complex emotional tasks with small samples, especially in modal alignment, feature fusion, and fine-grained emotion recognition. Significantly improved. This paper provides a systematic reference for the selection and optimization of pre-trained models for multi-modal sentiment analysis, all of which promote the improvement of the robustness and generalization of the model in the scenarios of small samples, social media, and complex emotions.

Keywords: Multimodal Sentiment Analysis, Social Media, Few-shot Learning, Modality Fusion, Pre-trained Models

1. Introduction

With the proliferation of social media and user-generated content, sentiment analysis has evolved from a text-only task into a multimodal discipline that incorporates various data sources such as audio, video, and images. Multimodal sentiment analysis (MSA) leverages complementary information from different modalities to better understand users' emotions and opinions. Compared to traditional text-based approaches, MSA provides richer sentiment cues, which are essential in handling ambiguity, sarcasm, or conflicting signals often found in natural communication.

MSA is particularly suitable for application scenarios such as customer service, mental health monitoring, content moderation, and human-computer interaction, where user communication often involves unstructured and multimodal signals. These domains require systems to interpret nuanced

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.PO24936

emotional expressions conveyed not only through text but also via tone, facial expressions, and visual context [1].

In recent years, the field has seen growing interest in MSA due to the advancement of powerful pre-trained models. Pre-trained language models such as BERT and RoBERTa have significantly improved natural language understanding. At the same time, models with cross-modal learning capabilities, such as CLIP and GPT-4, have demonstrated the potential to learn joint representations from both textual and visual data. As recent studies suggest, these foundation models are driving progress in affective computing and MSA research [2,3].

For instance, Amin et al [3]. conducted a wide evaluation of ChatGPT on affective computing benchmarks, highlighting both its strengths and limitations in emotion recognition tasks. Similarly, Debess et al [4]. explored the use of GPT-4 for sentiment analysis in low-resource languages, emphasizing the adaptability of general-purpose models to specialized affective tasks.

In addition to general-purpose large language models, several dedicated frameworks tailored for MSA have emerged. Yang et al [5]. proposed a contrastive learning-guided framework to align sentiment intensities across modalities, improving cross-modal consistency. Zhao et al [6]. introduced shared-private memory networks to extract both modality-specific and shared emotional features. These developments reflect a shift from task-specific architectures toward more scalable and transferable approaches based on foundation models.

Nevertheless, the field still faces several challenges, including modality imbalance, missing or noisy data, and the limited availability of large-scale, high-quality multimodal sentiment datasets.

This study aims to systematically review the role of pre-trained models in multimodal sentiment analysis. Specifically, it summarizes representative approaches based on foundation models, analyzes their performance on benchmark datasets, and discusses emerging techniques such as prompt tuning and contrastive learning. The remainder of this paper is organized as follows: Section 2 reviews five multimodal sentiment analysis pre-trained models focusing on their architectures, fusion strategies and performance; Section 3 presents a comparative analysis of these models; Section 4 conducts application case analysis; Section 5 discusses development and challenges; Section 6 concludes the paper.

2. Multimodal sentiment analysis technology

This section reviews five multimodal sentiment analysis pre-trained models proposed in recent years, focusing on their model architectures, fusion strategies, few-shot adaptability, and performance in social media scenarios. The design concepts, application contexts, and performance of these models are introduced and compared. The models discussed are the Multimodal Gating Transformer (MGT), SensoryT5, Adaptive Multimodal Prompt-Tuning (AMP), Multimodal Modality Enhancement Network (MMEN), and VLP2MSA.

2.1. MGT

MGT proposes a novel multimodal fusion mechanism that dynamically integrates non-linguistic modal data—such as vision and audio—into a Pre-trained Language Model (PLM), to address key challenges in traditional sentiment analysis models. These challenges include coarse-grained fusion methods, noise interference from non-linguistic modalities, and insufficient feature utilization due to distributional discrepancies across modalities. This approach enhances both the robustness and accuracy of sentiment analysis models.

DOI: 10.54254/2755-2721/2025.PO24936

MGT comprises four main components: an input module, PLM, a Multimodal Flow Module (which includes Cross-Modal Additive Attention and a Dynamic Gating Mechanism), and an output module. Anchored by the PLM with language as the primary modality, MGT incorporates multimodal features through Cross-Modal Additive Attention. Simultaneously, it employs an innovative Gating Mechanism to dynamically filter the integrated information [7].

Notably, to tackle the issue of modality alignment and mitigate the information loss typically associated with traditional contrastive learning, MGT introduces cross-sample matching loss and sample-specific margin loss to align distributional differences among modalities [7]. By fully leveraging the complementary strengths of linguistic and non-linguistic modalities, MGT achieves efficient multimodal data fusion while enhancing model robustness through dynamic noise filtering.

2.2. SensoryT5

Current sentiment analysis models primarily rely on semantic features for emotion recognition, which often leads to difficulties in distinguishing subtle emotions and identifying the nuanced relationships between emotions and sensory inputs. SensoryT5 addresses these limitations by innovatively integrating human multimodal sensory information into a text-processing framework, thereby enhancing the model's capability for fine-grained emotion classification.

SensoryT5 consists of four main modules: an input layer, the T5 backbone model, a sensory information transformation component, and a sensory attention mechanism. Its backbone is the T5 (Text-to-Text Transfer Transformer) model, which is based on an encoder-decoder architecture. SensoryT5 transforms textual input into word vectors, constructs corresponding sensory vectors using the Lancaster sensory dictionary, and then projects these vectors via two linear transformations to align with the dimensional space of T5 [8].

A sensory attention layer is incorporated based on the decoder output to generate fused representations of sensory and semantic information, ultimately producing emotion classification labels [8]. SensoryT5 demonstrates strong performance in fine-grained emotion differentiation and effectively simulates the human process of emotional perception and understanding. Minimizing information interference resulting from feature injection, helps preserve the semantic structure within the pre-trained model.

In addition, SensoryT5 exhibits broad adaptability across tasks and scenarios and is scalable to other model frameworks such as BERT, RoBERTa, and XLNet. Its emphasis on emotionally salient vocabulary further enhances its interpretability.

2.3. AMP

AMP introduces an adaptive multimodal prompt generation module that automatically generates prompts from both images and text to guide a PLM in accurately determining sentiment polarity. This approach addresses key challenges such as insufficient integration of visual and textual information and the oversight of modality-specific differences, thereby improving performance in few-shot learning scenarios.

The architecture of AMP comprises four main modules: an input module, a pre-trained multimodal feature extraction module, an adaptive multimodal prompt generation module, and a sentiment prediction module. AMP first extracts initial fused image-text features using the pre-trained multimodal feature extraction module. Learnable vectors are then introduced, concatenated, and passed through a Bidirectional Long Short-Term Memory (BiLSTM) network to construct contextually relevant prompt vectors [9].

DOI: 10.54254/2755-2721/2025.PO24936

Depending on the task type, aspect word information is selectively incorporated, and a multimodal prompt is generated via a Multi-Layer Perceptron (MLP). This generated prompt is then sequentially embedded at specific positions to predict the word distribution for the masked token [9].

The primary advantages of AMP lie in its ability to enhance the language model's comprehension and discrimination capabilities in joint image-text sentiment analysis through structural innovation and semantic bridging. Additionally, the use of prompt tuning mechanisms improves the model's adaptability under few-shot learning conditions.

2.4. MMEN

To address key challenges in multimodal sentiment analysis—such as the dominance of the text modality, insufficient inter-modality collaboration, and the neglect of weaker modalities like images and audio, which often lead to sentiment recognition bias and modality shift—MMEN proposes a Transformer-based Multimodal Modality Enhancement Network along with an Adaptive Contrastive Learning (ACL) method. This approach enhances inter-modality collaboration while preserving modality independence, and it increases the involvement of weaker modalities, thereby improving the model's capacity to recognize authentic emotions.

The architecture of MMEN consists of four main components: multimodal data augmentation, unimodal feature extraction, modality fusion, and ACL. MMEN extracts modality-specific information from text, audio, and visual data, converting them into feature vectors using a combination of pre-trained models and Transformer encoders. During the fusion process, modality embeddings and a learnable [CLS] token are incorporated to facilitate unified feature fusion across modalities [10].

Within this framework, the Multimodal Cooperative Contrastive Learning (MCCL) module strengthens consistency across modalities, while the Uncertainty Weighting (UW) mechanism dynamically adjusts each modality's contribution to the final prediction. After processing through a MLP, MMEN outputs both unimodal and multimodal sentiment predictions [10].

By introducing contrastive learning to maintain modality independence, MMEN enhances the model's decision-making capabilities. Furthermore, by improving the representational capacity of weaker modalities, it effectively mitigates the modality shift issue resulting from text modality dominance.

2.5. VLP2MSA

VLP2MSA extends Vision-Language Pre-training (VLP) to multimodal sentiment analysis in order to address several limitations in existing models, including the weak expressive capacity of the visual modality, suboptimal modality fusion architecture design, over-reliance on the text modality, and the underutilization of visual information. These issues often lead to degraded model performance when dealing with complex emotional expressions.

Unlike traditional frameworks, VLP2MSA introduces alignment and prompting modules prior to interaction, allowing different modalities to perform collaborative representation before fusion. This design significantly enhances fusion effectiveness. The architecture of VLP2MSA consists of four components: a text encoder, a video encoder (Inter-frame Hybrid Transformer, IHT), a Video-Text Contrastive (VTC) module, and a multimodal integration encoder [11].

VLP2MSA processes raw text and video frames through the text encoder and the IHT to extract modality-specific features, thereby improving the expressive power of the text representations.

DOI: 10.54254/2755-2721/2025.PO24936

These features are then aligned using contrastive learning in the VTC feature space. Finally, the fused features are passed through the multimodal integration encoder for emotion intensity prediction [11].

VLP2MSA offers several advantages, including end-to-end training, a clear and concise architecture, and reduced reliance on the audio modality. It provides a novel and effective approach for optimizing multimodal fusion in sentiment analysis tasks.

3. Comparative analysis and summary

To more intuitively present the characteristics of related research, this paper summarizes the five aforementioned models (MGT, SensoryT5, AMP, MMEN, and VLP2MSA) in a comparative table (Table 1). The models are analyzed across five dimensions: main problems addressed, datasets used, main baseline models, model performance comparison, and model characteristics.

All methods included in the table are designed for few-shot and social media scenarios. They primarily focus on enhancing fusion capabilities, improving fine-grained sentiment recognition, reducing learning difficulty under few-shot conditions, and optimizing the contributions of weaker modalities. The table highlights the practical potential and adaptability of these approaches, offering clear guidance for model selection and serving as a valuable reference for future research.

Table 1. Comparison of pretrained multimodal sentiment analysis models in few-shot learning and social media scenarios

Name	Main problems addressed	Datasets used	Main baseline models	Model performance comparison	Model characteristics
MGT	Improving the performance of pre-trained language models on tasks such as sentiment analysis remains an important research objective.	CMU-MOSI, CMU-MOSEI, MELD,	TFN, LMF, MISA, Self-MM, UniMSE, ITHP, ConFEDE	Achieved outstanding performance in multimodal sentiment analysis (MSA) and emotion recognition (MER) tasks, attaining state-of-the-art (SOTA) results on the CMU-MOSI, CMU-MOSEI, and MELD datasets.	Efficiently integrates multimodal information, enhances model robustness, and improves the effectiveness of multimodal fusion.
Sensory T5	Enhancing the fine-grained classification capability of sentiment analysis remains a key research objective.	GoEmotions, Empathetic Dialogues (EmD), ISEAR, Emolnt, SemEval 2018, Ghosh IAC-V2, SUBJ, PC (Comparative Sentences)	BERT, RoBERTa, T5, LLaMA3, HypEmo, SarDeCK	In the sentiment classification task, accuracy and F1 score improved by up to 3.0% compared to the baseline T5 model. The model also demonstrated strong performance in sarcasm classification.	Simulates human emotional cognition, enhances the understanding of complex emotions, and demonstrates strong cross-task applicability.

DOI: 10.54254/2755-2721/2025.PO24936

Table 1. (continued)

AMP	Addressing the scarcity of annotated multimodal data	Twitter15, Twitter17, MVSA-Single, MVSA-Multiple	BERT, RoBERTa, PVLM, UP-MPF, PVLM, A²II	Outperformed existing state-of-the-art methods across all four multimodal sentiment analysis datasets.	Significantly enhances the language model's ability to accurately analyze and utilize semantic information from images.
MMEN	Preserving the unique information of each modality after multimodal fusion	CH-SIMS v2.0	TFN, LMF, MulT, MISA, AV-MC	Improved Acc2 and F1 scores by 2.51% and 2.43%, respectively, compared to the AV-MC model, and enhanced weak sentiment classification accuracy by 2.69%.	Reduces the gap between unimodal and multimodal representations while preserving richer and more valuable information.
VLP2MSA	Enhancing inter-modal synergy and alleviating the low contribution of the visual modality in multimodal sentiment analysis	MOSI, MOSEI, CH-SIMS	TFN, MulT, Self-MM, MMIM, HyCon, MISA	By utilizing only two modalities, the model surpassed previous methods that relied on three modalities across three widely used datasets.	Improves the effectiveness of visual information and the accuracy of multimodal fusion.

4. Application case analysis

Multimodal Sentiment Analysis (MSA) substantially elevates the accuracy and robustness of sentiment analysis by leveraging complementary information from textual, auditory, visual, and physiological signals. With ongoing advancements in large language models (LLMs) and graph neural networks (GNNs), the application potential of MSA in scenarios such as cross-cultural emotion recognition and real-time emotion tracking will further expand.

In the field of multimodal sentiment analysis, Xianbing Wang et al [12]. proposed a Transformer-based Multimodal Enhancement Network (MMEN). This framework leverages multi-head attention mechanisms to fuse multimodal information, employs Modality Cooperative Contrastive Learning (MCCL) to align unimodal and multimodal features for enhanced fusion, and incorporates a weight dynamic adjustment mechanism (U) to optimize modality contributions. MMEN outperformed the AV-MC model by 2.51% on Acc2 and 2.43% on F1 metrics. Compared to AV-MC, it demonstrated superior performance across regression metrics, particularly achieving a 2.69% improvement on Acc2_weak.The effectiveness of this multimodal fusion strategy was validated on the CH-SIMS v2.0 dataset [13], where MMEN surpassed existing state-of-the-art methods across six key metrics:Acc2(%),Acc2_weak(%) ,F1_score(%) ,Corr ,R_square,MAE.

Guofeng Yi et al [14]. developed the Inter-frame Hybrid Transformer architecture (VLP2MSA), using raw video frames as input to capture spatiotemporal features, generating textual prompts from video content and aligning features through Video-Text Contrastive (VTC) loss. On MOSI, MOSEI and CH-SIMS datasets, VLP2MSA surpassed previous three-modality-dependent works using only video and text modalities, especially demonstrating the most outstanding performance on MOSI and MOSEI with over 4% improvement in both binary classification accuracy and F1-score. Subsequent optimizations enhanced visual modality importance through inter-frame hybrid transformer modules and instance-level enhanced fusion schemes, simplified cross-modal fusion processes, and comprehensively surpassed existing methods including TFN [15], LMF [16], and MFM [17].

DOI: 10.54254/2755-2721/2025.PO24936

Qingqing Zhao et al [18]. designed the Multimodal Gating Transformer (MGT), introducing cross-modal additive attention and gating mechanisms based on the T5 pre-trained model to fuse multimodal information centered on the language modality, aligning modal distributions through cross-modal boundary loss and matching loss, achieving advanced performance in Multimodal Sentiment Analysis (MSA) and Multimodal Emotion Recognition (MER) tasks; particularly in MSA tasks surpassing optimal models, SensoryT5 demonstrated significant performance improvements over Pre-trained Language Models (PLMs): compared to the best-performing T5 on GoEmotions and Empathetic Dialogues (EmD) datasets, SensoryT5 achieved increases in both accuracy and F1-score—in GoEmotions: accuracy improved by 1.3%, F1-score by 1.3%; in EmD: accuracy by 0.9%, F1-score by 1.1%; similarly outperforming the best-performing PLM RoBERTa on ISEAR and EmoInt datasets—in ISEAR: accuracy by 0.4%, F1-score by 0.4%; in EmoInt: both accuracy and F1-score by 1.1%.

5. Development and challenges

The future development directions of multimodal emotion analysis encompass technological innovation, application expansion, and interdisciplinary integration: Technologically, it is necessary to advance self-supervised and generative learning to enhance data utilization efficiency, develop dynamic and lightweight models to optimize real-time performance and edge deployment capabilities, and construct hierarchical modeling frameworks to strengthen complex semantic representation and reasoning; In terms of application expansion, efforts should focus on deepening emotion analysis in vertical fields such as healthcare and autonomous driving, as well as emotion interaction perception and scenario generation with specific emotional tones in metaverse virtual social interactions; In interdisciplinary integration, it can combine with cognitive science to analyze the social diffusion mechanisms of multimodal content based on psychological emotional dimensions and sociological communication models.

However, the field still faces several core challenges. 1. Differences in expression styles, information granularity, and semantic hierarchies across modalities (such as images, speech, and text) often lead to information loss or misinterpretation during data alignment. Additionally, due to the temporal characteristics of multimodal data, how to effectively achieve temporal synchronization and ensure the temporal consistency of information across modalities remains a critical issue to be solved. 2. Cross-modal semantic contradictions and temporal synchronization difficulties are core alignment challenges in multimodal emotion analysis caused by modal heterogeneity. The former stems from differences in expression styles, information granularity, and semantic hierarchies among modalities like images, speech, and text, leading to information loss or misinterpretation during fusion (e.g., polarity conflicts between visual "smiling" and textual "anger"). The latter faces dual challenges of physical-layer device delays (e.g., audio-visual asynchrony) and logical-layer emotional response asynchrony (e.g., language preceding facial expressions) due to the temporal nature of multimodal data, making it difficult to ensure temporal consistency. Together, these two issues restrict models from accurately understanding complex emotions. 3. Unstable few-shot learning caused by data scarcity is a core challenge in multimodal emotion analysis, arising from the contradiction between scarce annotated data and complex multimodal modeling. Specifically, insufficient sample sizes make models prone to overfitting, mistakenly treating superficial modal co-occurrences as universal laws and reducing emotion classification accuracy; high annotation costs lead to label noise problems, with noise rates in small sample sets potentially reaching 20%-25%, amplifying misjudgment risks; during cross-domain migration, modal distribution differences

DOI: 10.54254/2755-2721/2025.PO24936

degrade model generalization capabilities, causing significant performance declines in niche scenarios (e.g., low-resource language analysis), thus restricting practical applications.

6. Conclusion

This study systematically reviews the applications of pre-trained models in multimodal sentiment analysis (MSA), focusing on few-shot and social media scenarios, and analyzes the technical pathways, performance, and adaptive characteristics of representative models. Through comparative analysis of models such as MGT, SensoryT5, AMP, MMEN, and VLP2MSA, it is found that pre-trained models significantly improve model performance in emotional feature extraction, modal alignment, and few-shot generalization capabilities through innovations in cross-modal fusion mechanisms (e.g., MGT's gating mechanism and cross-sample matching loss), sensory semantic injection (e.g., SensoryT5's human perception simulation framework), prompt tuning strategies (e.g., AMP's adaptive prompt generation), and weak modality enhancement techniques (e.g., MMEN's adaptive contrastive learning and VLP2MSA's visual-language pretraining extension).

However, current research still faces three core challenges: Firstly, cross-modal semantic contradictions and temporal synchronization difficulties caused by modal heterogeneity, such as polarity conflicts between visual "smiling" and textual "anger," as well as physical audio-visual delays and asynchronous emotional responses, which restrict models from accurately parsing complex emotions; Secondly, unstable few-shot learning triggered by the scarcity of multimodal annotated data, manifested as high overfitting risk, label noise sensitivity, and insufficient cross-domain transfer capability, particularly evident in scenarios such as low-resource language emotion analysis; Thirdly, the contradiction between model interpretability and lightweight requirements—while existing complex architectures enhance performance, they struggle to meet the practical application needs of edge computing and real-time emotion tracking.

Future research can break through into three aspects: technology, applications, and interdisciplinary integration. Technically, it is necessary to explore self-supervised learning and generative models to enhance data utilization efficiency, construct dynamic lightweight architectures to optimize real-time deployment capabilities and develop hierarchical modeling frameworks that enhance the interpretability of emotional reasoning. In applications, efforts should deepen emotion analysis in healthcare, emotion perception in autonomous driving, and expand emotional interaction generation and scenario modeling in the metaverse. In interdisciplinary terms, integrating cognitive science and sociological theories can help analyze the social diffusion mechanisms of multimodal content, providing biologically inspired theoretical support for emotion computing. In summary, pre-trained models inject new momentum into multimodal emotion analysis, but their sustainable development relies on the joint advancement of technological innovation, application-driven development, and interdisciplinary integration to achieve more robust, interpretable, and scenario-adaptive emotional intelligence systems.

Authors contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] N. J. Shoumy, L.-M. Ang, K. P. Seng, D. M. Rahaman, and T. Zia, "Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals, " Journal of Network and Computer Applications, vol. 149, p. 102447, 2020.

DOI: 10.54254/2755-2721/2025.PO24936

[2] M. Amin, E. Cambria, and B. Schuller, “Will affective computing emerge from foundation models and general AI? A first evaluation on ChatGPT, ” Preprint at http: //arxiv.org/abs/2303.03186, 2023.
[3] M. M. Amin, R. Mao, E. Cambria, and B. W. Schuller, “A wide evaluation of ChatGPT on affective computing tasks, ” IEEE Transactions on Affective Computing, 2024.
[4] I. N. Debess, A. Simonsen, and H. Einarsson, “Good or bad news? exploring GPT-4 for sentiment analysis for Faroese on a public news corpora, ” Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 7814–7824.
[5] Y. Yang, X. Dong, and Y. Qiang, “CLGSI: A multimodal sentiment analysis framework based on contrastive learning guided by sentiment intensity, ” Findings of the Association for Computational Linguistics: NAACL 2024, pp. 2099–2110.
[6] X. Zhao, Y. Chen, S. Liu, and B. Tang, “Shared-private memory networks for multimodal sentiment analysis, ” IEEE Transactions on Affective Computing, vol. 14, no. 4, pp. 2889–2900, 2022.
[7] Mai, S., Zeng, Y., Xiong, A., & Hu, H. (2025). Injecting Multimodal Information into Pre-trained Language Model for Multimodal Sentiment Analysis. IEEE Transactions on Affective Computing.
[8] Zhao, Q., Xia, Y., Long, Y., Xu, G., & Wang, J. (2025). Leveraging sensory knowledge into text-to-text transfer transformer for enhanced emotion analysis. Information Processing & Management, 62(1), 103876.
[9] Xiang, Y., Zhang, A., Guo, J., & Huang, Y. (2025). Adaptive multimodal prompt-tuning model for few-shot multimodal sentiment analysis. International Journal of Machine Learning and Cybernetics, 1-13.
[10] Hu, Y., Huang, X., Wang, X., Lin, H., & Zhang, R. (2024). Transformer-based adaptive contrastive learning for multimodal sentiment analysis. Multimedia Tools and Applications, 1-18.
[11] Yi, G., Fan, C., Zhu, K., Lv, Z., Liang, S., Wen, Z., ... & Tao, J. (2024). Vlp2msa: expanding vision-language pre-training to multimodal sentiment analysis. Knowledge-Based Systems, 283, 111136.
[12] Hu, Y., Huang, X., Wang, X., Lin, H., & Zhang, R. (2024). Transformer-based adaptive contrastive learning for multimodal sentiment analysis. Multimedia Tools and Applications, 84, 1385–1402.https: //doi.org/10.1007/s11042-024-19152-5
[13] Liu Y, Yuan Z, Mao H et al (2022) Make acoustic and visual cues matter: Ch-sims v2. 0 dataset and av-mixup consistent module. In: Proceedings of the 2022 international conference on multimodal interaction, pp 247–25
[14] Yi, Guofeng, et al. "VLP2MSA: Expanding Vision-Language Pre-Training to Multimodal Sentiment Analysis." Knowledge-Based Systems, vol. 283, 2024, 111136. https: //doi.org/10.1016/j.knosys.2023.111136
[15] A. Zadeh, M. Chen, S. Poria, E. Cambria, L.-P. Morency, Tensor fusion network for multimodal sentiment analysis, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017.
[16] Z. Liu, Y. Shen, V.B. Lakshminarasimhan, P.P. Liang, A. Zadeh, L.-P. Morency, Efficient low-rank multimodal fusion with modality-specific factors, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018.
[17] Y.H.H. Tsai, P.P. Liang, A. Zadeh, L.P. Morency, R. Salakhutdinov, Learning factorized multimodal representations, 2018, arXiv preprint arXiv: 1806.06176.
[18] Zhao, Q., Xia, Y., Long, Y., Xu, G., & Wang, J. (2025). Leveraging sensory knowledge into Text-to-Text Transfer Transformer for enhanced emotion analysis. Information Processing and Management, 62, 103876. https: //doi.org/10.1016/j.ipm.2024.103876