Multi Modal Image Text Sentiment Analysis Method for Social Media

Zelin Wang *

Department of Computer Science and Technology, Taiyuan University of Science and Technology, Shanxi, China

* Corresponding Author Email: 202320060423@stu.tyust.edu.cn

Abstract. With the widespread application of social media, the multimodal fusion feature of "text + image" in user-generated content has become a dominant form. Therefore, sentiment analysis based on multimodal social media images and texts is an important research topic today. However, sentiment analysis faces three major challenges: enhancing image modality expression, bridging the cross-modal semantic gap, and improving the ability to model complex texts. This paper reviews five typical multimodal sentiment analysis methods proposed in the past two years. These methods are categorized into image enhancement, semantic alignment, and large model integration based on the types of problems they primarily address. A comparative analysis of their technical paths, performance, and application scenarios is conducted to understand the research focus of each type of model, identify their shortcomings and challenges, and propose a cross-category fusion technology solution by integrating existing model methods. An integrated framework of "universal semantic encoding + dynamic modality adaptation + cross-scenario calibration" is constructed to improve the generalization and robustness of models, providing directions for future research on multimodal social media image-text sentiment analysis.

Keywords: Multimodal Sentiment Analysis, Semantic Alignment, Synthetic Image Enhancement, Large Model Fusion, Cross-Category Technology Fusion.

1. Introduction

The widespread application of social media has become an indispensable part of modern society. The multimodal fusion feature of text combined with images has become a major form, and more and more people tend to express their opinions, comments, and emotions online through text and images [1]. Most social media posts, such as those on Twitter and Weibo, contain a combination of images and texts, which carry users' emotional tendencies and opinion expressions [2].

Multimodal Sentiment Analysis refers to the integration of multiple data modalities, using various sensors to acquire and process information, predict the intensity and polarity of human emotions [3], and finally comprehensively identify the expressed emotions to improve the comprehensiveness and accuracy of emotion classification. Therefore, sentiment analysis based on multimodal social media images and texts has become an important topic in the current research field [4], and its research results are of great value and significance in fields such as social media emotion monitoring [1], auxiliary diagnosis of mental health [5], and e-commerce review analysis [6]. For example, by analyzing the subtle emotional clues in Weibo users' images and texts, the ability to detect complex emotions can be improved [7], combining the emotional tendencies of social media texts and images can help identify potential depression risks [5].

In the past three years, numerous innovative models and methods have been proposed for multimodal social media image-text sentiment analysis: A deep network based on multi-level attention processes image-text modalities through a multi-level attention mechanism. It first uses channel and spatial attention to generate dual-attention visual features, enhancing the emotion region capture ability of Convolutional Neural Network (CNN), then associates image regions with text semantics through semantic attention, and combines self-attention to extract emotion-enriched multimodal features. Its performance on datasets such as Multi - View Sentiment Analysis (MVSA) and Flickr has been verified to be superior to traditional fusion methods [2]. The Weibo synthetic image fusion model, aiming at the informality and emotional ambiguity of Weibo texts, uses Stable

Diffusion to generate synthetic images matching the text, and combines contrastive learning to realize deep interaction between text and visual features, significantly improving the accuracy of subtle emotion recognition on datasets such as Social Media Processing (SMP2020) [7]. The entity knowledge-guided model, oriented to aspect-level sentiment analysis, introduces entity class embeddings to guide text feature learning, and uses image scene descriptions and adjective-noun pairs as entity knowledge to achieve entity-level feature alignment through cross-modal Transformer, achieving state-of-the-art performance on the Twitter dataset [8]. The aspect-guided progressive fusion model takes aspects as the hub, and optimizes modality alignment in stages through three-layer contrastive learning: first narrowing the image-text gap, then promoting interaction through double-layer cross-modal attention, and finally realizing feature fusion, verifying its robustness on datasets such as multimodal aspect - based sentiment analysis (MASAD) [9]. The social media depression detection model focuses on the application of text sentiment analysis in the field of mental health. By comparing Bidirectional Encoder Representations from Transformers (BERT), A Robustly Optimized BERT Pretraining Approach (RoBERTa), and Long Short-Term Memory (LSTM) models, it is found that the Transformer architecture (BERT with an accuracy of 99.9%) is better in context understanding and anti-overfitting, and can effectively identify implicit depression tendency signals in social media texts [5].

This paper aims to systematically and comprehensively sort out the core methods, technologies, and innovations of the models mentioned in the above five literatures, classify the models into three categories: image enhancement, semantic alignment, and large model integration according to the research perspectives, conduct comparative analysis of their technical paths, performance, and application scenarios, then summarize the common shortcomings and propose reasonable solutions, providing references for future research.

2. Overview of Mainstream Technologies

Sentiment analysis faces three major challenges: enhancing image modality expression, bridging the image-text semantic gap, and improving the ability to model complex texts. The research models investigated in this chapter address the above challenges effectively by optimizing image features through dual-attention mechanisms or generative models, strengthening semantic fusion based on entity knowledge and aspect hubs, and breaking through the limitations of traditional deep learning with pre-trained models such as BERT, thus promoting the development of multimodal sentiment analysis.

2.1. Image Enhancement Category

2.1.1 Deep Multi-Level Attentive Network Model

To address the problems that existing multimodal sentiment analysis methods are difficult to capture complex associations between images and texts, ignore channel information, and overlook the enhancement effect of image-level features on text emotion words, Yadav et al. proposed Deep Multi-Level Attentive network (DMLANet) [2]. This method generates dual-attention visual features through a visual attention module combining channel and spatial attention — as visualized in Fig. 1, which details the workflow from feature map (M) processing via max pooling (MP)、global average pooling (GAP), to channel - attention maps (Ac) and spatial - attention maps (As) generation. Then, it extracts text features related to visual features using semantic attention, and obtains emotion - rich multimodal features for classification through self - attention. Its innovation lies in the integration of multi-level attention mechanisms to capture fine-grained associations, which has been verified to be superior on four real datasets. However, its effect is limited on datasets with weak cross-modal associations, and it does not consider the robustness in scenarios with missing modalities.

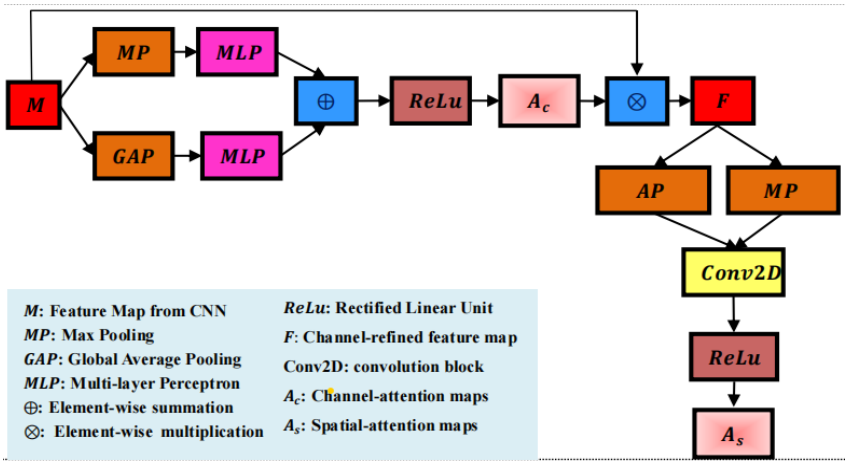

Fig 1. Block diagram explaining the Visual Attention Module [2]

2.1.2 Weibo Sentiment Analysis Model Integrating Text and Synthetic Images

Facing the difficulties in emotion capture caused by informal language, sarcasm, and missing context in Weibo texts, as well as the problems that existing models ignore visual clues and rely on single classification loss with limited generalization ability, Wang et al. proposed a multimodal method [7]. As shown in Fig. 2, this method processes Weibo posts through two parallel branches: the text branch uses BERT to extract text features, while the visual branch first generates synthetic images for sentiments via Stable Diffusion, then leverages ResNet - 50 (a typical CNN architecture) to extract image features. After fusing these modalities, it conducts training with cross - entropy loss and Information Noise Contrastive Estimation (InfoNCE) contrastive loss. Its innovation lies in supplementing visual information with synthetic images to mitigate text ambiguity, and enhancing subtle emotion detection through self - supervised contrastive learning — a design verified to achieve performance far exceeding traditional models [7].

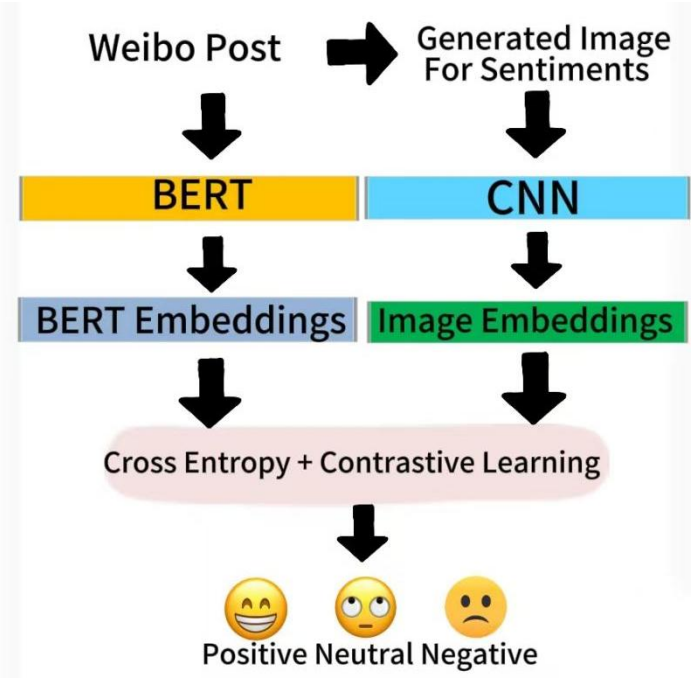

Fig 2. Framework of Multimodal Sentiment Analysis with Synthetic Visual Augmentation

2.2. Semantic Alignment Category

2.2.1 Entity Knowledge-Guided Image-Text Alignment Model

Xiang et al. addressed the difficulty in aligning image-text aspect features in joint multimodal aspect-level sentiment analysis, and proposed an entity knowledge-guided image-text alignment

network, pointing out that existing methods ignore entity-level semantic associations, leading to insufficient matching accuracy. Its principle uses entity knowledge as a cross-modal bridge: on the text side, entity class embeddings are introduced to guide the focus on entity semantics, and aspect features are strengthened with RoBERTa and self-attention, on the image side, adjective-noun pairs generated by DeepSentiBank and scene descriptions generated by Caption Transformer form entity knowledge. After extracting visual features through ResNet, the correlation is enhanced through interaction between cross-attention and entity knowledge. Multimodal fusion models associations through an interactive Transformer, outputs labels through conditional random fields, and optimizes training by combining text unimodal and multimodal losses. As shown in Fig. 3, the innovation lies in explicitly integrating entity knowledge into cross-modal alignment for the first time, guiding feature learning bidirectionally, and alleviating the semantic gap. The F1 scores on Twitter2015 and 2017 datasets reach 66.8% and 69.8% respectively, superior to the baseline [8].

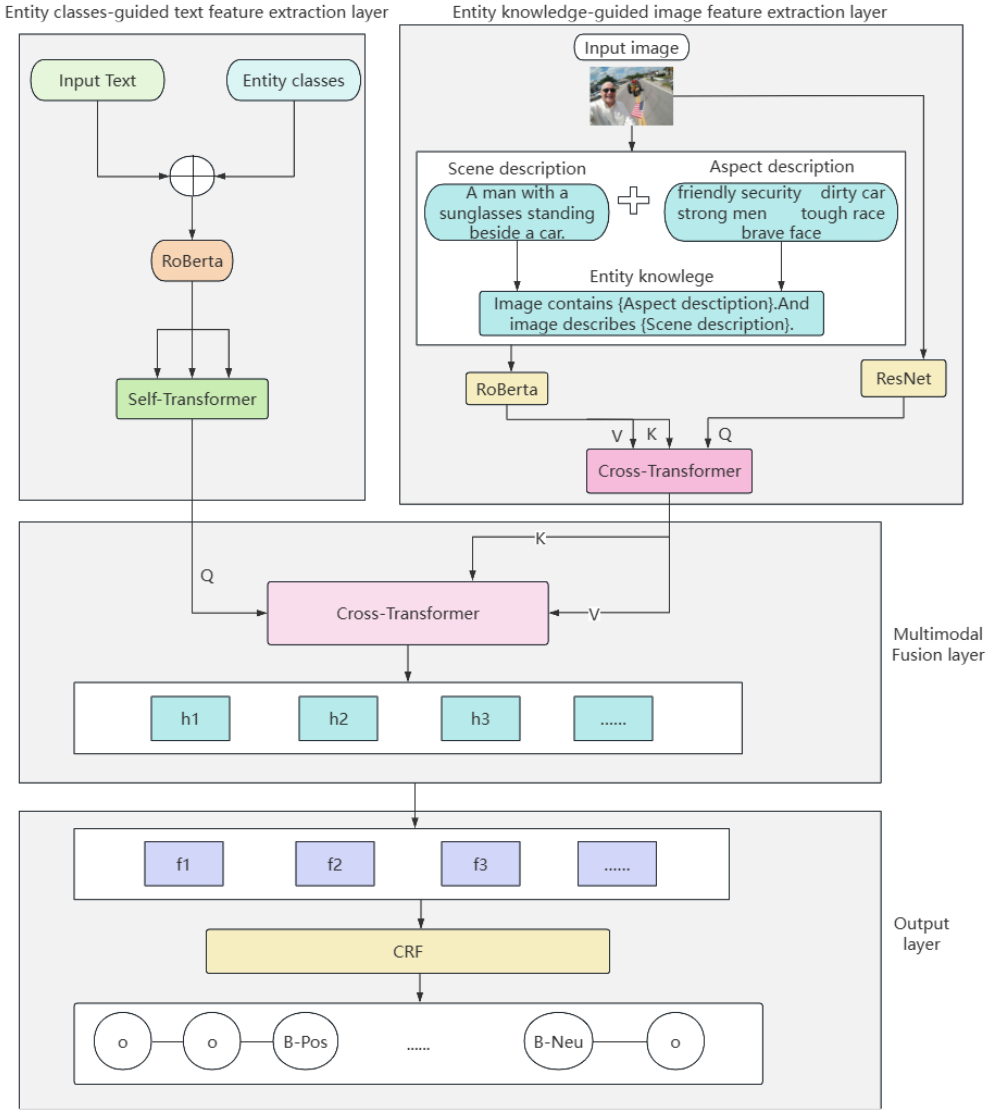

Fig 3. Two or more references Entity Knowledge - Guided Multimodal Alignment Network Architecture

2.2.2 Aspect-Guided Progressive Image-Text Fusion Model

Yan et al. addressed the problems of significant modality differences and insufficient cross-modal feature fusion in multimodal aspect-level sentiment analysis, and proposed an aspect-guided progressive image-text fusion framework. This study argues that existing methods mostly adopt a single fusion strategy, making it difficult to refine the alignment and integration of aspect-related features layer by layer [9].

Its technical principle takes "aspect" as the semantic hub and realizes progressive fusion through three-layer contrastive learning: the first layer is aspect-guided image-text contrastive learning, which narrows the distance between text-aspect and image-aspect in the semantic space based on aspect features, reducing the initial modality gap, the second layer is contrast-based cross-modal semantic interaction, which strengthens the association between aspects and modality features through a double-layer cross-attention mechanism (using aspects as queries to interact with text and image features respectively), and introduces contrastive loss to optimize alignment accuracy, the third layer is contrast-based multimodal semantic fusion, which uses mixup technology to achieve smooth transition of image-text features, retains modality specificity by combining self-attention and cross-attention, and finally encodes the fused features through Transformer, outputting emotion polarity through softmax. The total loss function integrates classification loss and three-layer contrastive loss to balance feature discriminability and alignment.

The innovation of this method lies in proposing a progressive alignment-fusion paradigm, which strengthens aspect relevance hierarchically from narrowing modality gaps to semantic interaction and then to feature fusion, furthermore, it retains unique modality information in fusion through symmetric mixup and bidirectional attention mechanisms. The Macro-F1 scores on Twitter-15, Twitter-17, and Multi-modal Aspect-based Sentiment Analysis Dataset (MASAD) datasets reach 74.31%, 70.02%, and 96.92% respectively, outperforming existing mainstream methods.

2.3. Large Model Integration Category

Isa et al. addressed the problems of insufficient application of Transformer models and deep learning models in large samples, weak generalization ability, and overfitting in existing depression detection research. Using 10,000 Twitter texts as the dataset, they compared the performance of BERT, RoBERTa, and LSTM models through sentiment analysis, its technical principle is to extract text semantic features based on sentiment analysis. As shown in Table 1, Table 2 and Table 3, BERT and RoBERTa are fine-tuned with pre-trained knowledge to adapt to sentiment classification, while LSTM captures sequence dependencies through a gating mechanism [5]. The method process includes data collection and exploration, preprocessing, segmentation and feature engineering, model training, and evaluation, the innovation lies in systematically comparing the performance of the three types of models on large samples and optimizing them with targeted nature, confirming the superiority of Transformer models [5].

Table 1. BERT Model Performance Across Three Epochs

Epoch	Training Loss	Validation Loss	Accuracy
1	1.3500%	0.5427%	99.9031%
2	0.0100%	0.4990%	99.9515%
3	0.4200%	0.4656%	99.9515%

Table 2. RoBERTa Model Performance Three Epochs

Epoch	Training Loss	Validation Loss	Accuracy
1	1.2000%	0.8503%	99.9031%
2	0.9100%	0.7821%	99.9031%
3	0.4500%	1.5193%	99.7092%

Table 3. LSTM Model Performance Across Ten Epochs

Epoch	Accuracy	Loss	Val Accuracy	Val Loss	Time per Epoch
1	34.62%	109.92%	31.00%	1.1009	12s
2	38.99%	107.86%	30.50%	1.1058	5s
3	86.53%	95.00%	32.5%	1.1438	6s
4	91.083%	57.56%	23.00%	1.3777	9s
5	96.54%	19.43%	29.50%	1.5883	6s
6	99.01%	06.39%	25.00%	1.8049	5s
7	99.75%	03.30%	27.00%	1.843	6s
8	99.66%	02.84%	30.00%	2.1162	9s
9	1.00%	00.82%	28.00%	2.2297	7s
10	99.87%	00.67%	25.00%	2.1965	5s

3. Comparison and Analysis

3.1. Comparison of Technical Paths

3.1.1 Summary of Core Features

Based on the core technical directions, the five methods are summarized as follows:

Table 4. Categories of Methods, Included Methods and Their Core Features

Category	Included Methods	Core Features
Image Enhancement	1.DMLANet 2. Enhancing Weibo Sentiment Analysis	Focusing on image feature enhancement or supplementation, strengthening visual features through attention mechanisms, or compensating for missing visual information with synthetic images.
Semantic Alignment	1. Entity Knowledge-Guided Method 2. Aspect-Guided Progressive Fusion Method	Focusing on semantic association modeling, guiding cross-modal semantic alignment and fusion through entity knowledge or aspect information.
Large Model Integration	Depression Detection Method	Relying on pre-trained large language models (BERT, RoBERTa) to process text, focusing on unimodal semantic encoding.

Table 4 systematically sorts out five types of core technical methods that are representative in the field of multimodal sentiment analysis and related areas. By clarifying the method categories, the specific included technologies, and their core features, it provides a clear framework for understanding the focus of different technical paths.

3.1.2 Differences in Attention Mechanisms

The image enhancement category includes DMLANet and the Weibo multimodal enhancement method. DMLANet adopts a multi-level attention mechanism: channel attention strengthens channels rich in image information, spatial attention focuses on emotion-related regions, semantic attention extracts emotional text features related to images, and self-attention filters redundancy to focus on key features, realizing fine modeling of visual features and cross-modal associations, the attention mechanism of the Weibo multimodal enhancement method is simpler: the text side relies on BERT's self-attention to capture contextual semantics, and the image side implicitly focuses on visual features through ResNet convolution layers, without complex cross-modal attention, mainly indirectly strengthening image-text associations through feature concatenation and contrastive learning.

The semantic alignment category includes the entity knowledge-guided method and the aspect-guided progressive fusion method. The entity knowledge-guided method designs cross-attention with entity knowledge as the hub: on the text side, entity class embeddings guide RoBERTa's self-attention to focus on entity-related text segments, on the image side, cross-attention based on entity knowledge extracts entity-related visual features, and then realizes semantic alignment of text and image entity features through cross-attention of the interactive Transformer, the aspect-guided progressive fusion method designs a double-layer cross-modal attention with "aspect" as the core: the first layer uses aspect features as queries to interact with image-text features to extract related features, the second layer deepens the association, and retains modality specificity through self-attention in the fusion stage to avoid semantic ambiguity.

The large model integration category, i.e., the depression detection method, has an attention mechanism limited to the self-attention of pre-trained models. BERT/RoBERTa capture depression-related semantics in texts through bidirectional self-attention, while LSTM captures sequence dependencies through a gating mechanism without explicit attention design, making it vulnerable to noise in long texts.

In general, the design of attention mechanisms is highly matched with their application scenarios: the image enhancement category focuses on visual feature strengthening and cross-modal association capture, the semantic alignment category emphasizes accurate alignment at the semantic level, and the large model integration category focuses on deep semantic mining of unimodal texts, providing diversified technical paths for sentiment analysis tasks under different needs.

3.2. Performance Comparison Analysis

Analysis from the Perspective of Core Performance Indicators in Table 5, the five methods show significant differences in their respective datasets, reflecting the close relationship between technical paths and task adaptability. The unimodal depression detection method, relying on pre-trained models such as BERT and RoBERTa, achieves an accuracy of up to 99.9% and 99.7% on the 10,000 Twitter text dataset, far exceeding the LSTM model in the same category, highlighting the advantages of pre-trained large language models in text semantic encoding. The image enhancement method DMLANet performs stably on multimodal datasets, with accuracies of 92.65% and 89.30% on GettyImages and Flickr datasets respectively, but slightly lower on MVSA series social media data, the Weibo multimodal enhancement method performs outstandingly on Chinese Weibo datasets, with accuracies of 99.10% and 96.22% on weibo_senti_100k and SMP2020 datasets respectively. Among semantic alignment methods, the entity knowledge-guided method achieves F1 scores of 66.8% and 69.8% on Twitter2015/2017, and the aspect-guided progressive fusion method achieves an average Macro-F1 of 96.92% on the cross-domain MASAD dataset, and 74.31% and 70.02% on Twitter-15/17, showing advantages in fine-grained analysis.

Table 5. Performance Metrics of Methods Across Different Categories, Names and Datasets

Method Category	Method Name	Dataset	Accuracy	F1 Score/Macro-F1
Image Enhancement	A Deep Multi-Level Attentive network (DMLANet)	MVSA-Single	79.47%	79.59%
		MVSA-Multiple	77.89%	75.26%
		Flickr	89.30%	89.19%
		GettyImages	92.65%	92.60%
	Enhancing Weibo Sentiment Analysis with Multi-Modal Learning	SMP2020 Weibo (Dataset1)	96.22%	96.00%
		weibo_senti_100k (Dataset2)	99.10%	98.95%
Semantic Alignment	Entity Knowledge-Guided Image-Text Alignment	Twitter2015	/	66.8%
		Twitter2017	/	69.8%
	Aspect-Guided Progressive Image-Text Fusion for Multimodal Aspect-Level Sentiment Analysis	Twitter-15	77.92%	74.31%
		Twitter-17	71.22%	70.02%
		MASAD	97.19%	96.92%
Large Model Integration	DEPRESSION DETECTION USING SENTIMENT ANALYSIS OF SOCIAL MEDIA	10,000 Twitter texts	99.9% (BERT), 99.7% (RoBERTa), 99.87% (LSTM)	/

One of the key reasons for performance differences is the adaptability to data characteristics. The Weibo multimodal enhancement method and the depression detection method are designed for specific platform data (Weibo, Twitter texts), which can effectively handle language characteristics such as slang and implicit expressions unique to the platforms, thus performing excellently. DMLANet achieves high accuracy on datasets with standardized texts such as GettyImages, benefiting from its fine modeling ability of associations between standardized texts and images, but has slightly weaker adaptability to more casual image-text content in social media.

Different technical paths also directly affect performance. Pre-trained models (BERT, RoBERTa) lead in accuracy in unimodal text analysis due to their strong context encoding ability, but lack cross-modal processing capabilities, image enhancement methods, by strengthening or supplementing visual features, perform better than text-only methods in image-text association scenarios, semantic alignment methods, focusing on fine-grained semantic alignment of entities and aspects, are irreplaceable in tasks of parsing specific emotional objects, but their overall indicators are lower than unimodal or general bimodal methods, which is closely related to the high complexity of fine-grained analysis tasks themselves.

3.3. Application Scenario Comparison

3.3.1 Social Media Emotion Detection

In the field of social media emotion detection, A Deep Multi-Level Attentive network (DMLANet) proposed by Yadav et al. plays an important role [2]. The network constructs a multi-level attention mechanism, which accurately captures the emotional associations of closely related image-text content through its unique design. Its visual attention module generates dual-attention visual maps along spatial and channel dimensions, thereby enhancing the expression ability of convolutional neural networks for image features. For example, it can highlight emotion key regions such as facial expressions and scene atmospheres in images, as well as related color and texture channel information. The semantic attention module models the semantic association between image regions and text words by extracting text features related to dual-attention visual features. For instance, when there is

a beautiful landscape in the image and the text mentions "intoxicating", it can effectively associate the emotions of the two. Finally, the self-attention mechanism automatically acquires emotion-rich multimodal features for classification. On platforms such as Flickr and Twitter, where data often present closely associated images and texts, DMLANet has shown excellent performance in tests on related datasets (such as Flickr and GettyImages), making it very suitable for overall emotion discrimination.

The Enhancing Weibo Sentiment Analysis method proposed by Wang et al. takes a different approach [7]. Considering that the Weibo platform is mostly text-dominated, often lacks visual information, and the text contains a large amount of informal language (such as internet buzzwords, abbreviations, and emoticons), this method supplements visual information by generating synthetic images. It uses advanced text-to-image generation technology to generate scene images corresponding to Weibo text content, providing more visual clues for understanding text emotions. Meanwhile, in feature extraction and fusion, it combines BERT to extract text contextual semantic features, CNN to extract visual features of synthetic images, and strengthens feature discriminability through contrastive learning, effectively improving the understanding of informal language in Weibo texts, and can more accurately judge the emotional tendency of Weibo content, especially suitable for text-dominated, visually missing scenarios such as Weibo.

3.3.2 Social Media Image-Text Content Review

The DMLANet proposed by Yadav et al. has significant advantages in social media image-text content review [2]. Thanks to its multi-level attention mechanism, it can focus on sensitive regions of non-compliant content with close image-text associations. At the visual level, spatial attention can quickly locate regions where non-compliant behaviors such as violence and pornography occur in images, and channel attention further strengthens the features of these regions, enabling reviewers to find problems more intuitively, at the semantic level, it associates sensitive words in texts with sensitive regions in images, avoiding misjudgments caused by single-modality understanding deviations, greatly improving review efficiency, and is suitable for image-text-rich social media platforms such as Instagram and Twitter, which can quickly screen out non-compliant content.

The Enhancing Weibo Sentiment Analysis method proposed by Wang et al. contributes significantly to content review on text-dominated platforms such as Weibo [7]. When dealing with non-compliant content with only text, this method restores potential scenarios through synthetic images. For example, for texts that implicitly describe non-compliant behaviors such as violence and terror, it generates corresponding image scenarios, combines multimodal features of text and synthetic images, more accurately judges whether the content is non-compliant, and effectively reduces misjudgments.

The entity knowledge-guided method proposed by Xiang et al. focuses on the review of non-compliant content containing entities [8]. This method uses entity knowledge, introduces entity class embeddings to guide the model to focus on specific entities in texts, such as people and events. At the same time, it uses image scene descriptions and aspect descriptions to strengthen the identification of associations between entities and non-compliant behaviors. When processing non-compliant content containing entities, it can reduce attention to irrelevant entities, greatly enhancing the pertinence of identifying entity-containing non-compliant content, and is suitable for social media platforms such as Weibo and Facebook where users often mention a large number of entities.

3.3.3 Mental Health Auxiliary Diagnosis

The depression detection method proposed by Isa et al. takes pre-trained models as the core, using powerful pre-trained language models such as BERT and RoBERTa for in-depth text understanding [5]. These models can accurately extract depression-related features, such as capturing implicit depression signals in texts such as "long-term insomnia", "no interest in anything", and "feeling life is meaningless". In tests on the Twitter depression dataset, the BERT model achieves an accuracy of 99.9%, and RoBERTa also reaches 99.7%, far exceeding the traditional LSTM model (prone to

overfitting), providing an efficient and reliable adaptation method for depression screening, and helping identify potential high-risk depression users in massive social media data.

The Enhancing Weibo Sentiment Analysis method proposed by Wang et al. also plays a unique role in mental health auxiliary diagnosis [7]. Weibo users often express negative emotions through implicit and informal language, and this method assists identification through synthetic images. When negative emotions are implicitly expressed in texts such as "I'm in a terrible mood today and don't want to talk", it generates corresponding dark and depressing scene images to supplement visual emotional clues. Moreover, through contrastive learning, it combines text features with synthetic image features to strengthen the capture ability of such weak and implicit emotional signals, helping identify whether users have mild depression or anxiety tendencies, and providing early warning for mental health intervention.

3.3.4 E-Commerce Review Analysis

The entity knowledge-guided method proposed by Xiang et al. takes entity knowledge as a key bridge in e-commerce review analysis [8]. When processing e-commerce reviews containing explicit entities, it focuses on product entities in reviews through entity class embeddings, such as "screen" and "battery" in mobile phone reviews. Combining scene descriptions of the entity in images (such as clear pictures displayed on the screen, the appearance of the battery, etc.) and aspect descriptions (such as "clarity" and "color performance" of the screen, "battery life" and "charging speed" of the battery, etc.), it closely associates emotions related to entities in texts and images, and accurately extracts emotional tendencies at the entity level. In tests on e-commerce-related datasets such as Twitter2015 and Twitter2017, the F1 values reach 66.8% and 69.8% respectively, making it very suitable for analyzing e-commerce reviews containing explicit product entities.

The aspect-guided progressive image-text fusion method proposed by Yan et al. adopts a hierarchical fusion strategy [9]. First, in the aspect guidance stage, it extracts various aspect words of products from texts, such as "cost performance", "appearance design", and "performance", which serve as the guide for subsequent fusion. In the progressive fusion stage, it first fuses aspect words in texts with local features of images, such as combining the "appearance design" aspect with image features of product appearance details, then gradually fuses global features to realize multi-domain aspect-level fine-grained analysis. This method can effectively distinguish emotional tendencies towards different aspects of products, such as judging that users are "satisfied with the appearance but think the performance needs improvement" for a product, providing detailed and targeted feedback information for merchants, and supporting them to make more accurate optimization decisions.

4. Challenges and Solutions

The common shortcomings of the above models and methods lie in weak semantic alignment capability across modalities or scenarios and insufficient robustness: DMLANet, the entity knowledge-guided method, and the aspect-guided progressive fusion method all rely on strong associations between image and text modalities. When modality inputs are incomplete or image-text semantic associations are loose, cross-modal feature alignment fails, and performance drops sharply, while the preprocessing model for depression detection and the text-dominated multimodal enhancement method (Weibo sentiment analysis) can handle pure text scenarios, but lack adaptability to cross-scenario semantic transfer, therefore, all methods have not effectively solved the "heterogeneous semantic gap", i.e., the unstable mapping relationship between the abstract semantics of text and the concrete features of images, leading to poor consistency in semantic understanding during cross-modal analysis.

The solution based on cross-category technology fusion is to construct an integrated framework of "universal semantic encoding + dynamic modality adaptation + cross-scenario calibration": integrating the universal semantic encoding of large models with multimodal fusion mechanisms, based on the strong semantic encoding capability of large models, embedding visual feature

enhancement technologies of the image enhancement category, and forming stable mappings through contrastive learning, combining fine-grained guidance of the semantic alignment category with dynamic modality adaptive strategies, adding "entity/aspect anchors" and dynamically adjusting modality weights through a gating mechanism, introducing cross-scenario knowledge distillation technology, distilling model knowledge from different scenarios into a universal model, and calibrating semantic deviations with scenario adaptation technologies, thereby improving the semantic consistency and robustness of cross-modal/cross-scenario analysis [10].

5. Conclusion

This paper summarizes relevant research on multimodal image-text sentiment analysis methods in recent years, classifying them into three categories: image enhancement, semantic alignment, and large model integration. Through comparative analysis, the applicability of various methods in scenarios such as social media emotion detection and mental health auxiliary diagnosis is clarified, and their common shortcomings are identified, such as weak cross-modal/cross-scenario semantic alignment capability and insufficient robustness. Finally, a cross-category technology fusion solution is proposed, constructing a framework of "universal semantic encoding + dynamic modality adaptation + cross-scenario calibration" to improve model generalization and robustness. In the future, it is necessary to further optimize the fusion mechanism, verify with more diverse data and scenarios, promote the more effective role of multimodal sentiment analysis in practical applications, and provide more accurate support for decision-making in related fields.

References

[1] Liu Y, Wang Z, Fang J, et al. Multimodal public opinion analysis based on image-text fusion. Journal of Computer Science and Exploration, 2022, 16(6): 1260-1274.

[2] Yadav A, Vishwakarma D K. A Deep Multi-Level Attentive network for Multimodal Sentiment Analysis. Proceedings of the 2023 ACM on Conference on Information and Knowledge Management, 2023: 1-11.

[3] Zhou H. The Application of Artificial Intelligence-based Multimodal Emotion Analysis. In: Wang Y (Ed.). Proceedings of the 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024), Advances in Intelligent Systems Research 185, 2024: 286-294.

[4] Zhao K, Zheng M, Li Q, et al. Multimodal Sentiment Analysis-A Comprehensive Survey From a Fusion Methods Perspective. IEEE Access, 2025, 13: 64556-64571.

[5] Isa Z O, Adewumi S E, Yemi-Peters V I. Depression Detection Using Sentiment Analysis of Social Media Text. FUW Trends in Science & Technology Journal, 2025, 10(1): 227-231.

[6] He X. Research on public attitude towards AI portrait generation technology based on social media big data and its application potential in e-commerce. E-Commerce Review, 2025, 14(4): 176-186.

[7] Wang C, Konpang J, Sirikham A, et al. Enhancing Weibo Sentiment Analysis with Multi-Modal Learning: Integrating Text and Synthesized Images with Contrastive Learning. IEEE Access, 2024, 11: 1-11.

[8] Xiang Y, Wu D, Cai Y, et al. Entity Knowledge-Guided Image-Text Alignment for Joint Multimodal Aspect-Based Sentiment Analysis. IEICE Transactions on Information and Systems, 2024, E107-A(1): 1-10.

[9] Yan Z D, Guo J J, Yu Z T. Aspect-Guided Progressive Fusion of Text and Image for Multimodal Aspect-Based Sentiment Analysis. Proceedings of the 23rd Chinese Computational Linguistics Conference, 2024: 454-466.

[10] Mao K B, Dai W, Guo Z H, et al. Review on the evolution and application of AI knowledge distillation technology. Journal of Agricultural Big Data, 2025, 7(2): 144-154.