Available Online: 10 June 2025 DOI: 10.54254/2977-3903/2025.23984

A review of multimodal aspect-based sentiment analysis

Tian'ang Chen

University of Electronic Science and Technology of China, Chengdu, China

chentianang@foxmail.com

Abstract. In the era of digital communication, the exponential growth of user-generated content across social media and online platforms has intensified the demand for effective emotion analysis tools. Traditional text-based sentiment analysis methods, however, often fall short in accurately capturing the nuances of human emotions due to their reliance on a single modality. Motivated by the need for more comprehensive and context-aware emotion recognition, this study systematically reviews the literature on both unimodal and multimodal aspect-level sentiment analysis. By comparing different approaches within the multimodal domain, we identify existing challenges and emerging trends in this research area. Our findings highlight the potential of integrating multiple modalities—such as text, images, and audio—to enhance the precision of sentiment detection and suggest future directions for advancing multimodal sentiment analysis.

Keywords: Aspect-Based Sentiment Analysis, Multimodal Aspect-Based Sentiment Analysis, Large Language Models (LLMs)

1. Introduction

In the era of digital communication, sentiment analysis has become an essential tool for understanding public opinion, customer satisfaction, and social trends. Traditional sentiment analysis primarily relies on textual data. However, with the increasing availability of multimedia content, Multimodal Sentiment Analysis (MSA) has emerged as a more comprehensive approach. MSA integrates information from multiple modalities—such as text, images, audio, and video—to provide a richer understanding of human emotions.

Aspect-level sentiment analysis has also garnered significant attention, with researchers shifting their focus from coarse-grained to fine-grained sentiment analysis. Within this domain, a substantial body of work has been dedicated to aspect-level sentiment analysis.

Despite the potential benefits of MSA, several challenges remain. Traditional unimodal sentiment analysis methods often fail to capture the full spectrum of human emotions due to their reliance on a single modality. In contrast, MSA leverages multiple modalities to provide more accurate sentiment predictions. However, effectively integrating and aligning information from different modalities remains a significant challenge. Early fusion methods, such as simple concatenation, often fail to capture the complex interactions between modalities. More advanced methods, including attention-based and tensor-based fusion, have shown promise but still face issues related to information redundancy and computational complexity.

This review aims to provide a comprehensive overview of the current landscape of MSA, with a focus on the comparison between unimodal and multimodal methods. By summarizing existing approaches and analyzing their strengths and weaknesses, we aim to identify gaps in the current research and suggest future directions for advancing MSA. This work is significant for both academic researchers and industry practitioners seeking to develop more effective sentiment analysis tools capable of handling the complexity of multimodal data.

The remainder of this paper is organized as follows: In Section 2, we provide a brief overview of existing aspect-based sentiment analysis methods, including unimodal and multimodal approaches. Section 3 compares unimodal and multimodal methods, with a focus on the fusion and alignment strategies used in MSA. Section 4 discusses future trends in Multimodal Aspect-Based Sentiment Analysis (MABSA), highlighting potential areas for further research. Finally, Section 5 summarizes the key findings and contributions of this review.

2. Related works

2.1. Aspect-Based Sentiment Analysis

Classical sentiment analysis was coarse-grained and could not meet the demand for fine-grained sentiment analysis. Therefore, researchers began to explore Aspect-Based Sentiment Analysis (ABSA). For instance, Abhijit Bhowmik et al. presented a groundbreaking method that employed sentiment analysis within an aspect-based context to decode the complex emotional subtleties embedded in student feedback. They classified sentiments into positive, negative, and neutral categories [1]. However, most ABSA datasets are relatively small and limited to these three sentiment categories. To address this issue, G. Kontonatsios et al. introduced Feedback ABSA (FABSA), a novel large-scale and multi-domain ABSA dataset comprising feedback reviews. FABSA was manually annotated using a hierarchical categorization framework that included seven parent and twelve child aspect categories [2].

With large language models (LLMs) demonstrating strong performance across many natural language processing tasks, some researchers have conducted ABSA studies based on LLMs. For example, L. Meng et al. introduced an innovative In-Context Learning (ICL) structure featuring a novel aspect-aware ICL example selection method to enhance the performance of LLMs in dimensional ABSA (dimABSA). Additionally, they investigated the performance of LLMs on the dimABSA task [3].

However, these studies primarily focus on unimodal (text-based) ABSA. In real-world applications, sentiment analysis often involves multiple modalities such as text, video, audio, and images. Due to the lack of analysis of non-textual modalities, existing unimodal ABSA methods are insufficient for such applications. Consequently, researchers have begun to explore multimodal ABSA.

2.2. Multimodal Aspect-Based Sentiment Analysis

Multimodal sentiment analysis is an emerging and dynamic field within natural language processing and machine learning. It focuses on the comprehensive interpretation of emotions conveyed through multiple modalities, including text, video, audio, and images. Unlike traditional sentiment analysis that relies solely on textual data, multimodal sentiment analysis aims to capture a more nuanced and holistic understanding of sentiment by integrating diverse sources of information. Existing research primarily addresses the challenges of representing, aligning, and fusing information from different modalities to create coherent and meaningful sentiment representations. Additionally, there is a growing emphasis on multitask learning, where models are designed to handle multiple objectives simultaneously, such as identifying sentiment while also detecting relevant aspects or features. This integration of multiple modalities and tasks not only enhances the accuracy and robustness of sentiment analysis but also provides a more comprehensive view of human expression in various contexts.

To introduce these methods more systematically, we divide them into five categories, with their features summarized in Table 1.

Table 1. Multimodal sentiment analysis: method categories

Category	Method	Features	References
Attention Mechanism-Based Methods	Visual-Text Fusion	Integrates visual and textual features with implicit attention on sentiment-related attributes	[4]
	Instruction Tuning	Uses large vision-language models to improve text-image fusion with a plug-and-play selector	[5]
	Spatio-Temporal Feature Enhancement	Enhances feature extraction from continuous image sequences using convolutional neural networks	[6]
	Aspect-Aware Attention Module (AOM)	Identifies text tokens and image blocks semantically aligned with aspects using graph convolutional networks	[7]
	Prompt Optimization	Combines fixed and learnable prompts with attention mechanisms to optimize the prompt encoder	[8]
Contrastive Learning-Based Methods	Supervised Angular Contrastive Learning	Captures generalized representations of each modality using triplet loss to enhance multimodal representations	[9]
	Modality Smoothing Fusion	Reduces modality gaps with a multi-channel attention mechanism for deep feature fusion	[10]

Table 1. Continued

Deep Learning-Based Methods	Four-Layer Model Architecture	Data layer, single-modality feature extraction layer, multimodal fusion layer, and sentiment analysis layer	[11]
	ConvLSTM	Replaces CNN pooling layers with LSTM to capture long-term dependencies	[12]
	BERT Models	Fine-tuning BERT models to enhance performance in sentiment analysis tasks	[13]
Feature Alignment and Enhancement Methods	Target-Oriented Feature Alignment (TOFA)	Enriches image features with target-related details to improve sentiment classification accuracy	[14]
	Adaptive Modality-Specific Weight Fusion Network (AdaMoW)	Analyzes dynamic relationships between modalities with a unimodal feature generator to retain modality-specific information	[15]
Other Methods	Hybrid Pattern	Combines fixed and learnable prompts with attention mechanisms for prompt optimization	[8]
	RNG Framework	Reduces multi-level modality noise and semantic gaps using global relevance, information bottleneck, and semantic consistency constraints	[16]
	Instruction Tuning	Uses large vision-language models to improve text-image fusion with a plug-and-play selector	[5]

2.2.1. Methods based on attention mechanisms for MABSA

Several methods based on attention mechanisms have been proposed. For instance, A. Aggrawal et al. integrated visual and textual features to predict sentiment in GIFs, incorporating attributes such as facial emotion detection and OCR-generated captions. Although not explicitly framed as such, these attributes can be regarded as implicit attention mechanisms to highlight sentiment-related features [4]. J. Feng et al. applied instruction tuning to the Multimodal Aspect-Based Sentiment Classification (MABSC) task, leveraging large vision-language models to improve text-image fusion. They introduced a plug-and-play selector to automatically choose the most relevant instruction, thereby reducing the impact of irrelevant image noise on sentiment classification results [5]. J. Bianbian et al. enhanced the attention weights of a Convolutional Neural Network (CNN) by leveraging traditional spatiotemporal feature keypoints, thereby improving feature extraction from continuous image sequences in the expression modality [6].

Since Graph Convolutional Networks (GCNs) often outperform CNNs in various tasks, attention mechanisms have been integrated with GCNs for MABSA. For example, R. Zhou et al. developed an Aspect-Aware Attention Module (AOM) to concurrently identify text tokens and image blocks semantically aligned with specific aspects. A GCN was employed to model interactions between vision and text, as well as within text, enabling precise aggregation of sentiment information [7].

Attention mechanisms have also been fused with LLMs. For instance, Z. Zhou et al. integrated fixed hand-crafted prompts with learnable prompts and used attention mechanisms to optimize the prompt encoder [8].

2.2.2. Methods based on contrastive learning for MABSA

Contrastive learning–based methods have been employed to merge different features from multimodal information. For example, C. D. Nguyen et al. proposed a Supervised Angular-based Contrastive Learning framework that uses a self-supervised triplet loss to capture generalized representations for each modality [9]. This enhances the discrimination and generalizability of multimodal representations and addresses biases in the fusion vector's modality. Similarly, Y. Xiang et al. introduced a modality-smoothing fusion strategy to reduce modality gaps and employed a multi-channel attention mechanism for deep feature fusion, thereby improving sentiment classification [10].

2.2.3. Methods based on deep learning for MABSA

Recent advancements in multimodal sentiment analysis have seen the development of various deep learning–based models to enhance multimodal integration and interpretation. One such model consists of four main components: the data layer, the single-modality feature extraction layer, the multimodal feature fusion layer, and the sentiment analysis layer [6]. Additionally, A. Hassan

and colleagues introduced ConvLSTM, a model that replaces the pooling layer in CNNs with an LSTM module to capture long-term dependencies and mitigate the loss of detailed local information in sequential data, achieving comparable performance with fewer parameters [12]. Moreover, Y. Wu et al. explored the structure and features of BERT models, demonstrating their strong capabilities in sentiment analysis tasks, especially after fine-tuning [13].

2.2.4. Methods based on feature alignment and enhancement for MABSA

D. Wang et al. integrated text information with both global coarse-grained and fine-grained image information for the MATE and MASC subtasks. The Target-Oriented Feature Alignment (TOFA) module enriches image features with target-related details, thereby enhancing sentiment classification accuracy [14].

Similarly, J. Zhang et al. introduced an Adaptive Modality-Specific Weight fusion network (AdaMoW) that analyzes dynamic relationships between modalities through data-driven techniques. A unimodal feature generator was designed to optimize multimodal fusion results by retaining more modality-specific information for sentiment analysis [15].

2.2.5. Other methods for MABSA

Z. Zhou et al. developed a hybrid pattern that integrates fixed hand-crafted prompts with learnable prompts and uses attention mechanisms to optimize the prompt encoder, enhancing the model's adaptability to diverse sentiment expressions [8].

Y. Liu et al. proposed the RNG framework for JMASA, which includes three constraints: the Global Relevance Constraint for instance-level noise reduction, the Information Bottleneck Constraint for feature-level noise reduction, and the Semantic Consistency Constraint for reducing multi-grained semantic gaps [16].

Additionally, J. Feng et al. applied instruction tuning to the MABSC task, leveraging large vision-language models to enhance the fusion of text and image modalities. A plug-and-play selector was introduced to automatically select the most relevant instruction, reducing the impact of irrelevant image noise [5].

In summary, various methods and techniques have been developed to integrate and interpret information from multiple modalities such as text, images, and video. These approaches include attention mechanisms that highlight sentiment-related features, contrastive learning methods for enhancing multimodal representations, deep learning models that capture long-term dependencies, feature alignment and enhancement techniques, and other innovative strategies such as hybrid patterns and instruction tuning. Collectively, these methods aim to address the challenges of multimodal integration and improve the accuracy and robustness of sentiment analysis.

3. Comparison of methods

3.1. Aspect-based sentiment analysis vs. Multimodal Aspect-Based Sentiment Analysis

To analyze the development trends in research, we compared Aspect-Based Sentiment Analysis (ABSA) and Multimodal Aspect-Based Sentiment Analysis (MABSA), as shown in Table 2:

Table 2. Comparison of methods

Comparison Dimension	Aspect-Based Sentiment Analysis (ABSA)	Multimodal Aspect-Based Sentiment Analysis (MABSA)
Data Modality Diversity	Primarily relies on text data to classify sentiments (positive, negative, neutral). Limited to textual information, often using text-only datasets.	Integrates multiple modalities (text, images, audio, and video) for a more comprehensive understanding.
Richness of Sentiment Information	Focuses on textual semantics, potentially missing non-verbal emotional cues.	Captures sentiment information from multiple modalities using techniques like contrastive learning.
Feature Fusion Mechanisms	Extracts features solely from text, optimizing text-based features. Focuses on improving sentiment classification from textual data.	Requires complex fusion mechanisms to integrate features from different modalities.
Model Complexity	Designed to optimize text-based sentiment analysis with fewer parameters.	More complex, handling multiple data types simultaneously.
Noise Handling	Can handle text noise but limited in addressing noise from other modalities. May struggle with irrelevant or misleading text data.	Introduces constraints to reduce noise from multiple modalities.
Depth of Semantic Information	Focuses on semantic information within the text, potentially overlooking emotional cues in other modalities.	Integrates semantic information from multiple modalities for a deeper understanding.

Table 2. Continued

Model Generalization	May perform well in specific domains but struggle with generalization across different contexts.	Better equipped to generalize across various domains and scenarios by leveraging diverse data types.
Sentiment Classification Accuracy	Achieves high accuracy within its domain but limited by reliance on a single modality.	Achieves higher accuracy through the integration of multiple modalities, leveraging complementary information.

3.1.1. Data modality diversity

Unimodal ABSA methods, such as those by Abhijit Bhowmik et al. and G. Kontonatsios et al., primarily rely on text data to classify sentiments into categories like positive, negative, and neutral [1, 2]. These methods focus exclusively on textual information, often using datasets limited to text reviews or feedback. In contrast, multimodal ABSA integrates multiple modalities, including text, video, audio, and images, to provide a more comprehensive understanding of sentiment. For instance, multimodal methods by R. Zhou et al. and J. Feng et al. leverage both visual and textual data to capture a richer set of emotional cues [5, 7].

3.1.2. Richness of sentiment information

While unimodal ABSA methods are effective for text-based sentiment analysis, they may overlook non-verbal cues carrying significant emotional information. For example, Wu et al. proposed BERT-based methods that focus on textual semantics but may miss emotional nuances present in visual or auditory data [13]. In contrast, multimodal ABSA methods, such as those using attention mechanisms or contrastive learning, capture sentiment information across multiple modalities. For example, Zhou et al. developed an aspect-oriented method that identifies semantically aligned text tokens and image blocks, offering a more nuanced understanding of sentiment [7].

3.1.3. Feature fusion mechanisms

Unimodal ABSA methods typically extract features solely from text, as utilized in most BERT-based models. These methods focus on optimizing text-based features for improved sentiment classification. In contrast, multimodal ABSA requires more complex fusion mechanisms to integrate features from different modalities. For example, Wang et al. used target-oriented feature alignment to enrich image features with target-related details, enhancing the overall sentiment analysis [14].

3.1.4. Model complexity

Unimodal ABSA models focus on improving the efficiency and accuracy of text processing. For instance, A. Hassan et al. designed a ConvLSTM-based model to optimize text-based sentiment analysis with fewer parameters [12]. Conversely, multimodal ABSA models are more complex as they must handle multiple data types simultaneously. For example, Zhang et al. introduced adaptive modality-specific weight fusion networks to analyze dynamic relationships between modalities [15].

3.1.5. Noise handling

Unimodal ABSA methods can manage text noise but are limited in addressing noise from other modalities. Traditional text-based methods may struggle with irrelevant or misleading text data.

Multimodal ABSA methods, however, employ techniques such as the Global Relevance Constraint and the Information Bottleneck Constraint to mitigate the impact of irrelevant data, thus enhancing robustness. For example, Liu et al. designed constraints to reduce noise from multiple modalities [16].

3.1.6. Depth of semantic information

Unimodal ABSA methods focus on semantic information within text, potentially missing emotional cues in other modalities. In contrast, multimodal ABSA integrates semantic information from multiple modalities, providing a deeper understanding of sentiment. For example, attention mechanisms are used to identify and weigh emotional cues from both text and visual data, offering a more comprehensive analysis.

3.1.7. Model generalization

Unimodal ABSA models often perform well in specific domains but struggle to generalize across different contexts. Text-based models trained on one type of review dataset may not perform effectively on another. Conversely, multimodal ABSA models, by

leveraging diverse data types, better generalize across various domains and scenarios. For example, models combining visual and textual data can adapt more easily to different types of reviews or feedback.

3.1.8. Sentiment classification accuracy

Unimodal ABSA methods achieve high accuracy within their domain but are constrained by reliance on a single modality. In contrast, multimodal ABSA methods, through the integration of multiple modalities, achieve higher accuracy in sentiment classification by leveraging complementary information from various sources.

3.2. Comparison of Multimodal Aspect-Based Sentiment Analysis methods

We also compare different multimodal aspect-based sentiment analysis methods, as shown in Table 3:

Table 3. Comparison of Multimodal Aspect-Based Sentiment Analysis methods

Category	Method	Key Characteristics	Comparative Advantage	References
Methods Based on Attention Mechanisms	Attention modules for text and images, combined with graph convolutional networks	Directly highlights sentiment-related features, easily integrated into existing models	High precision, low computational overhead	[5, 7]
Methods Based on Contrastive Learning	Self-supervised learning, triplet loss to reduce modality gaps	Learns generalized and robust representations, minimizes differences between modalities	Enhanced generalization, reduced modality gaps	[10]
Methods Based on Deep Learning	Deep neural networks, multi-layer architecture for fusion	Captures complex patterns and interactions, end-to-end training for task-specific optimization	Complex pattern recognition, end-to-end training	[6, 12]
Methods Based on Feature Alignment and Enhancement	Techniques to align and enrich features, modules for additional context or details	Ensures compatibility between features from different modalities, simple and efficient implementation	Simplicity and efficiency, feature compatibility	[14, 15]

3.2.1. Methods based on attention mechanisms

These methods leverage attention mechanisms to selectively focus on the most relevant parts of the input data, enhancing the model's ability to capture sentiment-related features across modalities.

1) Comparative advantage:

Precision over Generalization: Compared to contrastive learning methods, attention mechanisms offer a more direct approach to highlight sentiment-related features without relying on complex loss functions. For example, Zhou et al. developed an Aspect-Aware Attention Module (AOM) that concurrently identifies text tokens and image blocks semantically aligned with the aspects, enabling precise sentiment aggregation [7].

2) Efficiency over complexity: Unlike deep learning–based methods that may require extensive training, attention mechanisms can often be integrated into existing models with minimal overhead. For instance, Feng et al. applied instruction tuning to the MABSC task, using a plug-and-play selector to automatically choose the most relevant instruction, thereby reducing the impact of irrelevant image noise on classification outcomes [5].

3.2.2. Methods based on contrastive learning

These methods use contrastive learning to generate representations that can discriminate between different modalities and enhance model generalizability.

Key characteristics:

Employ self-supervised learning to capture generalized representations for each modality.

Use triplet loss or other contrastive losses to reduce modality gaps and improve feature fusion.

Comparative advantage:

Robustness over Precision: Compared to attention mechanisms, contrastive learning offers a more systematic approach to addressing modality biases and improving representation learning. For example, Nguyen et al. proposed a supervised angular-based contrastive learning framework using triplet loss to enhance the discrimination and generalizability of multimodal representations [9].

Structured Learning over Flexibility: Unlike feature alignment methods, contrastive learning does not require explicit feature alignment but learns to discriminate between modalities, as demonstrated by Xiang et al [10].

3.2.3. Methods based on deep learning

These methods rely on deep learning architectures to extract and fuse features from multiple modalities, often employing complex multi-layer models.

Key characteristics:

Utilize deep neural networks with multiple layers for feature extraction and fusion.

Often include components such as CNNs, LSTMs, and transformers.

Comparative advantage:

Comprehensive Integration over Precision: Compared to attention mechanisms, deep learning models capture more complex patterns and interactions between modalities. For example, the four-layer model provides a structured approach to integrating multiple modalities [6].

Performance over flexibility: Unlike contrastive learning, deep learning models do not rely on self-supervised learning and can be trained end-to-end. This allows for more direct task-specific optimization, as shown by Hassan et al. with the ConvLSTM model, which replaces the pooling layer in CNNs with LSTM modules to capture long-term dependencies while preserving local information [12].

4. Methods based on feature alignment and enhancement

These methods focus on aligning and enhancing features from different modalities to improve the accuracy and relevance of sentiment analysis.

Key characteristics:

Utilize techniques such as feature alignment and enhancement to ensure that features from different modalities are compatible and carry relevant sentiment information.

Often incorporate modules designed to enrich features with additional context or detailed information.

Comparative advantage:

Efficiency over Complexity: Compared to deep learning methods, feature alignment techniques are often simpler and require fewer computational resources. For example, Wang et al. introduced the Target-Oriented Feature Alignment (TOFA) module, which enriches image features with target-related details, thereby enhancing target-specific emotional content [14].

Structured Alignment over flexibility: Unlike attention mechanisms, feature alignment provides a more structured approach to ensuring compatibility between features from different modalities. This structure is particularly beneficial in tasks where feature compatibility is crucial for accurate sentiment classification, as demonstrated by Zhang et al. with the Adaptive Modality-Specific Weight Fusion Network (AdaMoW) [15].

Each type of multimodal sentiment analysis method has its own strengths and is suited to different types of tasks. Attention mechanisms offer precision and flexibility, making them ideal for tasks that require detailed sentiment analysis. Contrastive learning enhances generalizability and reduces modality biases, making it suitable for tasks involving diverse modalities. Deep learning methods provide comprehensive integration and high performance but often require extensive training and computational resources. Feature alignment and enhancement techniques ensure feature compatibility and relevance, making them particularly efficient for tasks involving complex multimodal data. By understanding the comparative advantages of each method, researchers can select the most appropriate approach for their specific multimodal sentiment analysis objectives.

5. Future trends

5.1. Better low-resource models needed

The primary challenge in low-resource sentiment analysis is the insufficiency of training data. For example, the PanoSent dataset [17], which combines multilingual and multimodal elements, has demonstrated the potential for sentiment analysis in multilingual contexts, providing new ideas for sentiment analysis in low-resource languages. In my opinion, several approaches can address this problem, as outlined below:

(1) Data augmentation and lexical enhancement: Future research is likely to focus more on data augmentation and lexical enhancement techniques. For example, training data can be expanded through back-translation and synonym substitution. Additionally, lexical enhancement frameworks based on multilingual models can effectively expand rare words in low-resource languages, thereby improving the model's understanding of sentiment-related vocabulary.

(2) Transfer Learning and Fine-Tuning of Pre-Trained Models: Transfer learning and fine-tuning pre-trained models are effective methods for addressing low-resource sentiment analysis. Future research may further explore how to optimize pre-trained models to better adapt to sentiment analysis tasks in low-resource languages. For instance, pre-training using multilingual models such as mBERT [18], followed by fine-tuning for specific low-resource languages, can significantly enhance performance.

(3) utilization of multimodal data complementarity: The complementarity of multimodal data will be increasingly utilized in low-resource environments. Combining textual, audio, and visual modalities can compensate for the limitations of single-modality data. For example, a multimodal framework that integrates visual and textual features—including facial emotion detection and OCR-generated captions—has demonstrated the value of multimodal data for enriching semantic information in sentiment analysis. Future research may explore additional multimodal fusion strategies to enhance low-resource sentiment analysis.

(4) Aspect-Enhanced Pre-Training: To enhance the model's sensitivity to aspect terms in multimodal data, aspect-enhanced pre-training tasks should be constructed using positive and negative samples based on aspects. For instance, the AESAL model [11] provides a valuable reference for low-resource sentiment analysis, particularly when multimodal data are scarce.

5.2. Joint models for multiple subtasks

5.2.1. Multitask learning and cross-modal fusion

Joint models for multiple subtasks will improve sentiment analysis performance by incorporating multitask learning and cross-modal fusion. For example, a cascaded cross-modal Transformer structure that fuses textual, audio, and visual modalities in sequence can simulate the way humans process multimodal information.

The RNG framework offers new perspectives on multitask learning [16]. It integrates global relevance constraints, information bottleneck constraints, and semantic consistency constraints to reduce instance-level and feature-level noise, as well as multi-grained semantic gaps, while optimizing the alignment between modalities.

5.2.2. Fine-grained sentiment analysis

Future sentiment analysis research will increasingly focus on recognizing fine-grained sentiment states, rather than simply categorizing sentiments into positive, negative, or neutral. Joint models for multiple subtasks can uncover more nuanced sentiment states—such as anxiety, excitement, or shame—through multimodal data analysis.

Aspect-based sentiment analysis methods, which traditionally classify sentiments into broad categories, have already demonstrated the potential for finer-grained analysis. Future research can build on this foundation to further expand sentiment categories and improve analysis precision.

5.2.3. Temporal modeling and dynamic sentiment analysis

Sentiment is a dynamically evolving process. Future research will place greater emphasis on the temporal modeling of sentiment. By introducing temporal models such as LSTM or Transformer architectures, the evolution of sentiment over time can be captured [12].

Temporal modeling shows particular promise in multimodal sentiment analysis. Techniques such as using spatiotemporal feature keypoints to extract expression modality features from continuous image sequences, coupled with enhanced attention weights in convolutional neural networks, can provide technical support for dynamic sentiment analysis.

5.3. Enhanced methods based on LLMs

Large Language Models (LLMs) have demonstrated significant advantages in multimodal sentiment analysis due to their ability to leverage rich semantic representations and effectively integrate information from multiple modalities. For example, models like BERT and its variants provide deep semantic representations that capture linguistic nuances, which is crucial for sentiment analysis. Additionally, LLMs can integrate information from multiple modalities more effectively. For instance, J. Feng et al. applied instruction tuning to the MABSC task, using large vision-language models to improve text-image fusion [5]. This approach leverages the strengths of LLMs to manage multimodal data, reducing the impact of irrelevant image noise on sentiment classification results.

Future developments in LLMs will significantly advance multimodal sentiment analysis by enhancing multimodal capabilities, specialization, and efficiency. These models will integrate text, images, audio, and video to provide richer context and more accurate sentiment classification. Furthermore, LLM-enhanced methods—such as model quantization and pruning—will make these powerful tools more accessible and sustainable for real-world applications.

6. Conclusion

This paper provides a comprehensive review and analysis of the current state of Multimodal Aspect-Based Sentiment Analysis (MABSA). It compares the differences between unimodal and multimodal ABSA approaches, examines alignment methods for multimodal information, and evaluates various fusion techniques for integrating data from different modalities.

Based on these analyses, this paper concludes that future trends in the field include the development of better low-resource models, the creation of joint models for multiple subtasks, and the adoption of enhanced methods based on Large Language Models. The insights and comparisons presented are intended to serve as a valuable reference for researchers initiating work in this important and rapidly evolving area.

References

[1] Bhowmik, A., Nur, N. M., Miah, M. S. U., & Karmekar, D. (2023). Aspect-based Sentiment Analysis Model for Evaluating Teachers' Performance from Students' Feedback. *AIUB Journal of Science and Engineering,* 22(3), 287–294. https://doi.org/10.53799/AJSE.V22I3.921

[2] Kontonatsios, G., Clive, J., Harrison, G., Metcalfe, T., Sliwiak, P., Tahir, H., &Ghose, A. (2023). FABSA: An aspect-based sentiment analysis dataset of user reviews. *Neurocomputing,* 562, 0–9. https://doi.org/10.1016/j.neucom.2023.126867

[3] Meng, L., Zhao, T., & Song, D. (2024). DS-Group at SIGHAN-2024 dimABSA Task: Constructing In-context Learning Structure for Dimensional Aspect-Based Sentiment Analysis. *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing*, 127–132. https://aclanthology.org/2024.sighan-1.15

[4] Aggrawal, A., & Varshney, D. (2024). Multimodal Sentiment Analysis: Perceived vs Induced Sentiments. *2024 Silicon Valley Cybersecurity Conference (SVCC)*. https://doi.org/10.1109/SVCC61185.2024.10637377

[5] Feng, J., Lin, M., Shang, L., & Gao, X. (2024). Autonomous Aspect-Image Instruction A2II: Q-Former Guided Multimodal Sentiment Classification. *2024 Joint International Conference on Computational Linguistics and Language Resources Evaluation (LREC) - Main Conference Proceedings*, 1996–2005.

[6] Bianbian, J., Rajamanickam, L., Lohgheswary, N., & Nopiah, Z. M. (2023). Multimodal Sentimental Analysis Based on Deep Learning. *Section A-Research Paper Eur.* (12), (5), 3567–3573. 10.48047/ecb/2023.12.si5a.0249.

[7] Zhou, R., Guo, W., Liu, X., Yu, S., Zhang, Y., & Yuan, X. (2023). AoM: Detecting Aspect-oriented Information for Multimodal Aspect-Based Sentiment Analysis. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, (1), 8184–8196. https://doi.org/10.18653/v1/2023.findings-acl.519

[8] Zhou, Z., Feng, H., Qiao, B., Wu, G., & Han, D. (2023). Syntax-aware Hybrid prompt model for Few-shot multi-modal sentiment analysis. *ArXiv, abs/2306.01312*.

[9] Nguyen, C. D., Nguyen, T., Vu, D. A., & Tuan, L. A. (2023). Improving Multimodal Sentiment Analysis: Supervised Angular Margin-based Contrastive Learning for Enhanced Fusion Representation. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 14714–14724. https://doi.org/10.18653/v1/2023.findings-emnlp.980

[10] Xiang, Y., Cai, Y., & Guo, J. (2023). MSFNet: modality smoothing fusion network for multimodal aspect-based sentiment analysis. *Frontiers in Physics, 11*(5), 1–10. https://doi.org/10.3389/fphy.2023.1187503

[11] Zhu, L., Sun, H., Gao, Q., Yi, T., & He, L. (2024). Joint Multimodal Aspect Sentiment Analysis with Aspect Enhancement and Syntactic Adaptive Learning. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence* (pp. 6678–6686). https://doi.org/10.24963/ijcai.2024/738

[12] Hassan, A., & Mahmood, A. (2017). Deep Learning approach for sentiment analysis of short texts. *2017 3rd International Conference on Control, Automation and Robotics (ICCAR)*, 705–710. https://doi.org/10.1109/ICCAR.2017.7942788

[13] Wu, Y., Jin, Z., Shi, C., Liang, P., & Zhan, T. (2024). Research on the application of deep learning-based BERT model in sentiment analysis. *Applied Computing Engineering,* 71(1), 14–20. https://doi.org/10.54254/2755-2721/71/2024ma

[14] Wang, D., He, Y., Liang, X., Tian, Y., Li, S., & Zhao, L. (2024). TMFN: A Target-oriented Multi-grained Fusion Network for End-to-end Aspect-based Multimodal Sentiment Analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16187–16197, Torino, Italia. ELRA and ICCL.

[15] Zhang, J., Wu, X., & Huang, C. (2023). AdaMoW: Multimodal Sentiment Analysis Based on Adaptive Modality-Specific Weight Fusion Network. *IEEE Access,* 11(April), 48410–48420. https://doi.org/10.1109/ACCESS.2023.3276932

[16] Liu, Y., Zhou, Y., Li, Z., Zhang, J., Shang, Y., &Zhang, C. (2024). RNG: Reducing Multi-level Noise and Multi-grained Semantic Gap for Joint Multimodal Aspect-Sentiment Analysis. *Proceedings - IEEE International Conference on Multimedia and Expo*. https://doi.org/10.1109/ICME57554.2024.10687372

[17] Luo, M., Fei, H., Li, B., Wu, S., Liu, Q., Poria, S., Cambria, E., Lee, M., & Hsu, Y. (2024). PanoSent: A Panoptic Sextuple Extraction Benchmark for Multimodal Conversational Aspect-based Sentiment Analysis. In Proceedings of *the 32nd ACM International Conference on Multimedia (MM '24).* Association for Computing Machinery, New York, NY, USA, 7667–7676. https://doi.org/10.1145/3664647.3680705

[18] Ye, J., Zhou, J., Tian, J., Wang, R., Zhang, Q., Gui, T., & Huang, X. (2023). RethinkingTMSC: An Empirical Study for Target-Oriented Multimodal Sentiment Classification. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 270–277. https://doi.org/10.18653/v1/2023.findings-emnlp.21