Trustworthy Multimodal Aspect-Based Sentiment Analysis with Vision-Language Models: Aspect-Guided Reasoning, Evidence Grounding, and Explainable Sentiment Understanding

Dr. Premnarayan Arya[1], Dr. (Prof.) Sashi Kant Gupta[2]
[1,2]Lincoln University College, Petaling Jaya, Selangor-47301, Malaysia
Email ID: pdf.Premnarayan@lincoln.edu.my, shashigupta@lincoln.edu.my

Abstract: Multimodal Aspect-Based Sentiment Analysis (MABSA) has emerged as a critical research area that aims to identify aspect-specific opinions from multimodal data such as text and images. Traditional sentiment analysis methods often fail to capture fine-grained opinions associated with individual aspects and struggle to leverage complementary visual information. Recent advances in Vision-Language Models (VLMs) and Multimodal Large Language Models (MLLMs), including CLIP, BLIP, LLaVA, and GPT-4V-inspired frameworks, have significantly enhanced multimodal understanding through cross-modal reasoning and knowledge transfer. Despite remarkable performance improvements, challenges related to trustworthiness, explainability, visual grounding, hallucination, fairness, and robustness remain unresolved. This review provides a comprehensive analysis of recent developments in trustworthy MABSA, emphasizing aspect-guided reasoning, evidence grounding, explainable sentiment understanding, and emerging Vision-Language architectures. Furthermore, we discuss benchmark datasets, evaluation methodologies, open challenges, and future research opportunities. The review highlights how grounded multimodal reasoning can establish reliable and transparent sentiment analysis systems suitable for real-world applications in e-commerce, healthcare, tourism, social media monitoring, and decision intelligence.

Keywords: Multimodal Aspect-Based Sentiment Analysis; Vision-Language Models; Explainable AI; Trustworthy AI; Evidence Grounding; Large Language Models; Sentiment Analysis.

I. Introduction

Sentiment analysis has become an essential component of Natural Language Processing (NLP), helping organizations understand opinions expressed through user-generated content. Traditional sentiment classification operates at document or sentence levels and frequently overlooks sentiment variations toward specific aspects of a product, service, or event. To address this issue, Aspect-Based Sentiment Analysis (ABSA) emerged as a fine-grained sentiment analysis approach capable of identifying sentiments associated with individual aspects.

The rapid growth of multimedia platforms such as Twitter/X, Instagram, Facebook, YouTube, and e-commerce websites has transformed online reviews into multimodal content consisting of text, images, videos, and emojis. Consequently, Multimodal Aspect-Based Sentiment Analysis (MABSA) has gained significant attention because visual content often provides complementary or even contradictory information that cannot be inferred from text alone. Recent surveys have highlighted MABSA as one of the fastest-growing subfields within sentiment analysis research [1], [2].

The emergence of Vision-Language Models (VLMs) such as CLIP and Multimodal Large Language Models (MLLMs) such as LLaVA has revolutionized multimodal understanding by learning joint

representations of images and text through large-scale pretraining. These models enable zero-shot reasoning, visual-textual alignment, evidence retrieval, and explainable decision-making capabilities beyond conventional deep learning architectures.

Although current MABSA systems have achieved remarkable performance, several critical challenges persist:

- Lack of transparency and explainability.
- Hallucinated sentiment explanations.
- Weak visual grounding.
- Dataset and demographic biases.
- Poor robustness under adversarial settings.
- Limited trustworthiness in high-stakes applications.

Therefore, developing trustworthy multimodal sentiment analysis systems has become an important research priority.

II. Evolution of Multimodal Aspect-Based Sentiment Analysis

A. From Text-Based ABSA to MABSA

Early ABSA systems relied on handcrafted features, machine learning techniques, and linguistic rules. With the advent of deep learning, recurrent neural networks, convolutional networks, and attention-based mechanisms significantly improved sentiment prediction accuracy.

As social media evolved, researchers recognized that images often carry sentiment cues unavailable in textual descriptions [3].

For example:

> ***Text:*** *"The hotel room was okay."*
> ***Image:*** *Displays a luxurious suite.*

In such situations, image information substantially influences sentiment interpretation. The first generation of MABSA systems used independent visual and textual encoders, followed by feature fusion layers. Despite their effectiveness, these approaches lacked semantic alignment between modalities [4].

B. Transformer-Based MABSA

The introduction of Transformer architectures dramatically improved multimodal learning. Cross-attention mechanisms allowed models to establish fine-grained interactions between image regions and textual aspects.

Transformer-based approaches introduced:

- Co-attention networks.
- Graph neural reasoning.
- Region-aware attention.
- Cross-modal alignment modules.

These models improved aspect-level sentiment classification but remained highly task-specific and computationally expensive [5], [6].

C. Vision-Language Models

Vision-Language Models represent a paradigm shift from architecture-centric design toward foundation models.

Examples: CLIP, BLIP, BLIP-2, Flamingo, LLaVA, GPT-4V-inspired systems.

CLIP demonstrated that large-scale vision-language pretraining creates transferable multimodal representations capable of supporting various downstream applications, including sentiment understanding. Studies have shown superior generalization performance compared with conventional visual sentiment models [7].

Similarly, advanced multimodal architectures such as LLaVA integrate visual understanding with large language reasoning capabilities, creating opportunities for aspect-level sentiment explanation and grounding.

III. Trustworthy Multimodal Sentiment Analysis

A. Definition of Trustworthiness

Trustworthy AI refers to systems that operate reliably, transparently, fairly, and safely.
In the context of MABSA, trustworthiness consists of:

1) Reliability
Models should provide stable predictions under varying inputs [8].

2) Explainability
Predictions should include understandable explanations.

3) Transparency
The reasoning process behind sentiment inference should be visible.

4) Fairness
Models must avoid demographic and cultural biases.

5) Robustness
Systems should resist adversarial attacks and noisy multimodal inputs.

B. Challenges

1) Hallucination
Large multimodal models occasionally generate unsupported explanations [9], [10]. For instance, a model may attribute negative sentiment to an object not present in the image.

2) Aspect Ambiguity
Images often contain multiple entities, making aspect localization difficult.

3) Cross-Modal Conflicts
Textual and visual information may convey contradictory sentiments.

4) Cultural Biases
Different populations interpret visual cues differently.

These issues directly impact explainability and trustworthiness. Recent grounding-focused models have been proposed to mitigate such problems.

IV. Aspect-Guided Reasoning in Vision-Language Models

Aspect-guided reasoning refers to explicitly incorporating aspect information into the sentiment decision-making process.

Instead of directly predicting sentiment, modern VLMs perform multi-step reasoning:

Step 1: Aspect Extraction

Example: Aspect = “food quality”

Step 2: Multimodal Evidence Collection

Text: “The burger tasted amazing.”
Image: Displays appetizing food.

Step 3: Cross-Modal Verification

The model verifies whether visual evidence supports textual claims.

Step 4: Sentiment Inference

Positive sentiment toward food quality.

Step 5: Explanation Generation

“The sentiment is positive because the review explicitly praises the taste, and the image visually depicts a well-presented meal.”

Recent explainable MABSA frameworks transform sentiment analysis into a generative reasoning task, enabling models to simultaneously classify sentiment and generate aspect-specific rationales [13]. Dependency-guided prompting further improves aspect discrimination and reasoning quality [12].

V. Evidence Grounding and Explainable Sentiment Understanding

Evidence grounding constitutes the foundation of trustworthy multimodal reasoning.

A. Textual Grounding

Textual grounding identifies sentiment-bearing phrases.

Example: Aspect: Battery
Evidence: "Battery drains within four hours."
Sentiment: Negative

B. Visual Grounding

Visual grounding associates sentiment decisions with image regions.

For instance:

- Aspect: Camera Quality
- Grounded Evidence: High-resolution photograph shown in uploaded image

Grounding techniques answer the crucial question: **"What evidence supports this prediction?"**

LLaVA-Grounding represents a major advancement in grounded visual conversation by linking generated responses to visual regions through object localization and segmentation mechanisms [14].

C. Explainable Sentiment Rationales

Explainable AI approaches generate human-readable rationales.
Typical explanation:
"The sentiment toward the restaurant ambiance is positive because the review mentions comfortable seating and the image shows an elegant interior with decorative lighting."

Advantages include:

- Increased user trust.
- Easier debugging.
- Regulatory compliance.
- Improved model accountability.

Recent studies report that MLLM-generated explanations can significantly improve human understanding of sentiment predictions when grounded in multimodal evidence.

VI. Benchmark Datasets and Evaluation Metrics

A. Public Datasets

Frequently used datasets include:

- Twitter-2015
- Twitter-2017
- MVSA-Single
- MVSA-Multiple
- HFM Dataset

These datasets remain the primary benchmarks for evaluating multimodal aspect-level sentiment classification methods.

B. Evaluation Metrics

1) Classification Metrics

- Accuracy
- Precision
- Recall
- F1-Score

2) Explainability Metrics

- Faithfulness
- Completeness
- Sufficiency

3) Grounding Metrics

- IoU (Intersection over Union)
- Localization Accuracy

4) Trustworthiness Metrics

- Calibration Error
- Fairness Score
- Robustness Measurement

Future evaluations should jointly assess classification performance, reasoning quality, explanation faithfulness, and grounding accuracy.

VII. Future Research Directions

Several promising research directions remain open.

A. Grounded Chain-of-Thought Sentiment Analysis

Future models should generate step-by-step multimodal reasoning traces supported by verifiable evidence.

B. Retrieval-Augmented MABSA

External knowledge bases can improve sentiment understanding for domain-specific reviews [15].

C. Causal Multimodal Sentiment Learning

Causal reasoning may enable models to distinguish genuine sentiment evidence from spurious correlations.

D. Explainable Multimodal Agents

Future MLLM-based sentiment agents could interactively justify their predictions.

E. Fair and Responsible AI

Researchers must develop techniques that reduce demographic, cultural, and geographic biases [11].

F. *Hallucination Detection*

Automatic identification of unsupported explanations remains an important open problem.

VIII. Conclusion

Trustworthy Multimodal Aspect-Based Sentiment Analysis represents a rapidly evolving research domain at the intersection of sentiment analysis, multimodal learning, explainable AI, and Vision-Language Modeling. Emerging VLMs and MLLMs such as CLIP and LLaVA have significantly advanced multimodal reasoning capabilities while creating opportunities for evidence-grounded and explainable sentiment understanding.

However, trustworthiness remains an unresolved challenge due to hallucinations, visual grounding limitations, bias, and transparency concerns. Future research should emphasize explainability, evidence grounding, causal reasoning, and human-centered evaluation methodologies. The integration of trustworthy AI principles with multimodal foundation models is expected to define the next generation of sentiment analysis systems capable of supporting real-world decision-making applications.

References

[1] H. Zhao, M. Yang, X. Bai, and H. Liu, "A survey on multimodal aspect-based sentiment analysis," IEEE Access, vol. 12, pp. 12039–12052, 2024.

[2] H. Yang, Y. Zhao, Y. Wu, S. Wang, T. Zheng, H. Zhang, Z. Ma, W. Che, S. Wang, and S. Wei, "Large language models meet text-centric multimodal sentiment analysis: A survey," Science China Information Sciences, vol. 68, no. 10, pp. 1–29, 2025.
[3] H. Liu, L. He, J. Liang, Z. Ren, H. Bi, and F. Li, "Dependency structure augmented contextual scoping framework for multimodal aspect-based sentiment analysis," arXiv preprint arXiv:2504.11331, 2025.
[4] H. Liu, L. He, and J. Liang, "Joint modal circular complementary attention for multimodal aspect-based sentiment analysis," in Proc. IEEE Int. Conf. Multimedia and Expo Workshops (ICMEW), Niagara Falls, ON, Canada, 2024, pp. 1–6.
[5] Z. Li, J. Yu, J. Yang, W. Wang, L. Yang, and R. Xia, "Generative multimodal data augmentation for low-resource multimodal named entity recognition," in Proc. 32nd ACM Int. Conf. Multimedia (MM), Melbourne, Australia, 2024, pp. 7336–7345.
[6] X. Liu, R. Li, S. Ye, G. Zhang, and X. Wang, "Multimodal aspect-based sentiment analysis under conditional relation," in Proc. 31st Int. Conf. Computational Linguistics (COLING), Abu Dhabi, UAE, 2025, pp. 313–323.
[7] S. Song, "Exploring large language models for multimodal sentiment analysis: Challenges, benchmarks, and future directions," arXiv preprint arXiv:2411.15408, 2024.
[8] M. Vatsa, A. Jain, and R. Singh, "Adventures of trustworthy vision-language models: A survey," in Proc. AAAI Conf. Artificial Intelligence, vol. 38, no. 20, 2024, pp. 22650–22658.
[9] Z. Bai, P. Wang, T. Xiao, T. He, Z. Han, Z. Zhang, and M. Z. Shou, "Hallucination of multimodal large language models: A survey," arXiv preprint arXiv:2404.18930, 2024.
[10] P. Sahoo, P. Meharia, A. Ghosh, S. Saha, V. Jain, and A. Chadha, "A comprehensive survey of hallucination in large language, image, video and audio foundation models," in Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, FL, USA, 2024, pp. 11709–11724.
[11] Y. Luo, M. Shi, M. O. Khan, M. M. Afzal, H. Huang, S. Yuan, Y. Tian, L. Song, A. Kouhana, T. Elze, Y. Fang, and M. Wang, "FairCLIP: Harnessing fairness in vision-language learning," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12289–12301.
[12] Z. Wang, Y. Tian, H. Wang, and Y. Song, "Explainable multimodal aspect-based sentiment analysis with dependency-guided large language model," arXiv preprint arXiv:2601.06848, 2026.
[13] J. Cao, J. Li, Z. Yang, and R. Zhou, "Enhanced multimodal aspect-based sentiment analysis by LLM-generated rationales," in Proc. Int. Conf. Neural Information Processing (ICONIP 2024), Lecture Notes in Computer Science, vol. 15294, Springer, 2025, pp. 1–12.
[14] H. Zhang, H. Li, F. Li, T. Ren, X. Zou, S. Liu, S. Huang, J. Gao, L. Zhang, C. Li, and J. Yang, "LLaVA-Grounding: Grounded visual chat with large multimodal models," in Proc. European Conf. Computer Vision (ECCV), 2024, pp. 19–35.
[15] S. M. Rafiuddin, S. Kamal, M. Rakib, A. Bagavathi, and A. Sen, "AdaptiSent: Context-aware adaptive attention for multimodal aspect-based sentiment analysis," arXiv preprint arXiv:2507.12695, 2025.