Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 47 No.10
oct 2026

Sentiment Analysis with Multiple Modalities in the Age of Generative AI using Multilevel Fusion

Sagar Patil[1], K Satyanarayan Reddy[2], Sandeep Sutar[3] and Suchita Patil[4]

[1, 2, 3] Department of Computer Science & Engineering, Srinivas University, Mukka, Mangaluru, Karnataka, India.

[1] Department of Computer Science & Engineering, BVCOE, Kolhapur, Maharashtra, India.

[3] Department of Computer Science & Engineering, ADCET, Kolhapur, Maharashtra, India.

[4] Department of Computer Science & Engineering, KITCOE, Kolhapur, Maharashtra, India.

Abstract

Introduction: Sentiment analysis is moving from predominantly text-centered classification toward multimodal understanding in which language, acoustic behavior, visual evidence, contextual metadata, and generated semantic descriptions jointly influence interpretation. Generative AI and large language models (LLMs) add a further layer of capability by producing contextual representations, synthetic training signals, semantic summaries, and explanations.

Objectives: This paper develops a conceptual framework for multilevel fusion that separates local modality evidence, intra-modal context, pair wise cross-modal interactions, global multimodal representations, and LLM-guided semantic reasoning. It also identifies design requirements for robustness, interpretability, data efficiency, and responsible deployment.

Methods: A structured narrative synthesis of recent journal literature is combined with architectural analysis. The paper organizes fusion strategies into five levels and maps Generative AI functions to each level. A reference architecture, comparative taxonomy, implementation workflow, and evaluation framework are proposed.

Results: The analysis indicates that multilevel fusion can reduce the information loss associated with single-stage fusion by preserving fine-grained modality cues while progressively modeling cross-modal dependencies. LLM-guided semantic supervision can provide a higher-level context signal, while generative augmentation can address sparse or imbalanced multimodal data. These benefits are accompanied by risks related to hallucination, synthetic-label inconsistency, computational cost, privacy, and domain shift.

Conclusions: A GenAI-era sentiment system should therefore be designed as a layered evidence-integration pipeline rather than as a single fusion operation. Multilevel fusion, uncertainty-aware modality weighting, controlled generation, and explanation auditing provide a coherent direction for future multimodal sentiment systems.

Keywords: Sentiment Analysis, Multimodal Sentiment Analysis, Generative Artificial Intelligence, Large Language Models, Multilevel Fusion.

1. Introduction

Sentiment analysis (SA) seeks to infer affective orientation from human-generated content. Classical approaches were largely text centric, but contemporary communication frequently combines written language with speech, facial behavior, images, video, emojis, and situational context. Multimodal sentiment analysis (MSA) therefore treats sentiment as a distributed signal rather than as a property contained in one channel. Reviews of the field consistently identify fusion as a central technical problem because heterogeneous modalities differ in sampling rate, representation, reliability, and semantic granularity (Gandhi et al., 2023; Zhu et al., 2023).

The emergence of transformer architectures and pretrained language models has expanded the capacity of sentiment systems to represent long-range linguistic dependencies. More recently, Generative AI and LLMs have introduced capabilities that are not limited to classification. A generative model can summarize context, produce auxiliary descriptions, create candidate labels, translate between representational spaces, and generate natural-language rationales. Recent literature has begun to connect LLMs with multimodal fusion, including LLM-guided feature enhancement and reasoning-oriented pipelines (Kim & Bartolini, 2026; Zhao et al., 2025).

Recent reviews characterize MSA as a rapidly developing field spanning text, image, audio, video, multimodal fusion, and affective computing (Lai et al., 2023; Pandey & Vishwakarma, 2024; Do et al., 2024). End-to-end multimodal affect recognition subsequently showed the value of jointly modeling raw or minimally processed text (Rupapara et al., 2021), audio, and visual information with attention mechanisms (Tzirakis et al., 2020).

Despite these advances, simply inserting an LLM into a multimodal pipeline does not solve the fusion problem. A model may still discard shallow but sentiment-relevant cues, over-weight one modality, amplify spurious correlations, or treat generated text as ground truth without sufficient verification. Recent research on multi-level fusion, hierarchical interaction, missing modalities, and causality-aware fusion suggests that effective MSA requires explicit management of information at several representational depths (Wang et al., 2024; Huang et al., 2025; Liu et al., 2024).

This paper uses the term multilevel fusion to describe a staged architecture in which information is integrated at more than one representational depth. The proposed view distinguishes local feature fusion, intra-modal contextualization, pair wise cross-modal fusion, global multimodal integration, and semantic reasoning. The objective is not to claim a single universally optimal architecture, but to provide a coherent design framework for researchers developing GenAI-enabled sentiment systems.

2. Objectives

The paper focuses on the sentiment analysis on multimodal data with application of latest technological advancement in generative AI and fusion methods which are explored using the following objectives.

• To conceptualize multilevel fusion architecture that integrates textual, acoustic, visual, and contextual evidence without prematurely collapsing modality-specific information.

• To examine how Generative AI and LLMs can support semantic enrichment controlled augmentation, cross-modal reasoning, and explanation generation within a fusion pipeline.

• To identify evaluation dimensions covering predictive quality, robustness to missing or noisy modalities, calibration, interpretability, computational efficiency, and generation reliability.

• To formulate implementation and governance recommendations for responsible deployment of GenAI-enhanced sentiment analysis.

3. Background and Related Work

The evolution of MSA can be viewed as a progression from feature concatenation toward interaction-aware representation learning. Early fusion combines modality features before prediction, whereas late fusion combines modality-specific predictions. Hybrid and hierarchical designs introduce intermediate interaction stages. Attention-based methods further allow the model to assign context-dependent importance to modality features. Surveys show that no single fusion family dominates every setting because the appropriate strategy depends on the task, modality completeness, dataset scale, and required latency (Gandhi et al., 2023; Zhu et al., 2023).

Recent research increasingly focuses on hierarchical and multi-level representations. Wang et al. (2024) proposed cross-modal hierarchical fusion with explicit treatment of conflicts between modalities. Li et al. (2024) introduced multi-level textual-visual alignment for multimodal aspect-based sentiment analysis. Zhu et al. (2025) addressed information dilution through modality decomposition, while a 2025 cross-modal multi-layer feature-fusion study showed the value of combining shallow and deep representations through adaptive gating. These studies motivate retaining multiple levels of evidence rather than relying only on final-layer embeddings.

More recent work has explored vision-language pre-training, text-centric hierarchical fusion, cross-modal attention, causality-aware fusion, and multi-scale representation learning (Yi et al., 2024; Hou et al., 2024; Huang et al., 2025; Li et al., 2026). Generative AI and LLMs introduce another layer of capability. Rather than treating the language model solely as a classifier, a research system can use an LLM for semantic enrichment, weak supervision, synthetic example generation, explanation, prompt-based adaptation, and multimodal reasoning. Journal evidence indicates growing interest in LLM-based sentiment analysis and multimodal LLMs, while also highlighting the need for grounding, evaluation, and careful comparison with specialized discriminative systems (Krugmann & Hartmann, 2024; Yang et al., 2025; Mu et al., 2025).

Another important direction is robustness to incomplete data. Real-world multimodal streams frequently contain missing audio, blurred images, unavailable video frames, or text without visual context. Modality-translation and multimodal-foundation-model approaches have therefore explored ways to infer or compensate for missing information (Liu et al., 2024; Zhao et al., 2025). A robust system should distinguish observed evidence from generated or reconstructed evidence so that downstream reasoning does not treat an inferred modality as equivalent to a directly observed signal.

Generative AI changes the role of auxiliary information. Instead of using generation only for synthetic samples, a model can produce semantic captions, contextual summaries, pseudo-labels, or explanations. LLM-guided sentiment work has demonstrated the feasibility of combining generated contextual information with encoder representations, while current multimodal foundation models increasingly support joint image-text reasoning. However, generated content can contain hallucinated details and therefore requires confidence estimation, provenance tracking, and consistency checks (Kim & Bartolini, 2026; Zhao et al., 2025).

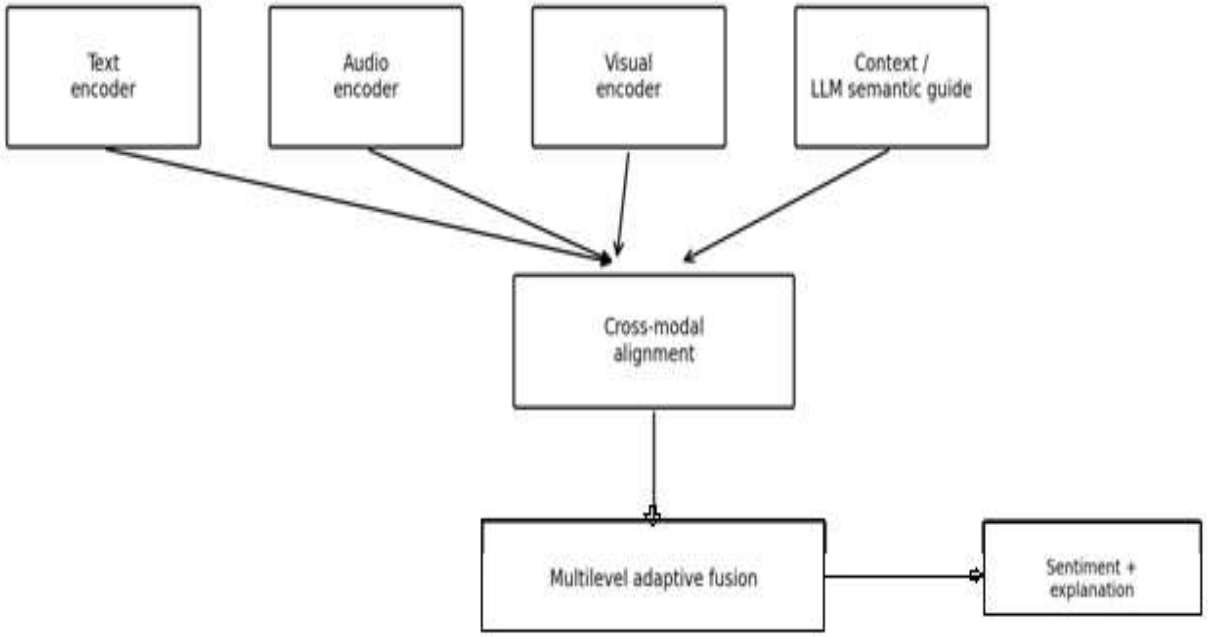

Figure1. Conceptual multilevel fusion architecture using GenAI for sentiment analysis

Table I. Proposed taxonomy of multilevel fusion

Fusion level	Primary evidence	Typical operation	Main benefit
Local	Tokens, acoustic frames, visual patches	Encoder-level feature extraction	Preserves fine-grained cues
Contextual	Within-modality sequences	Self-attention / temporal modeling	Captures local-to-global context
Cross-modal	Text–audio, text–vision, audio–vision	Cross-attention / interaction blocks	Models complementary and conflicting cues
Global	Joint multimodal representation	Adaptive or gated fusion	Balances modalities for prediction
Semantic	LLM-generated context and rationale	Reasoning, verification, explanation	Adds higher-level interpretation

4. Proposed Multilevel Fusion Framework

The proposed framework treats sentiment prediction as a sequence of evidence transformations. Let text, audio, and visual inputs be represented by X_t, X_a, and X_v, respectively. Contextual information is denoted by X_c. Each modality is first encoded independently so that its characteristic structure is preserved. The resulting representations are then aligned in a shared latent space before progressively interacting across levels. A semantic guide generated by an LLM is introduced as an auxiliary signal rather than an unquestioned label source.

At Level 1, modality-specific encoders produce local representations. Text may be encoded using a transformer language model, audio using a speech or acoustic encoder, and visual data using a vision transformer or convolutional backbone. The design principle is to preserve temporal, lexical, prosodic, and spatial evidence. At Level 2, self-attention or temporal aggregation captures within-modality context. This separation is important because some sentiment signals become meaningful only after local features are contextualized.

At Level 3, pair wise cross-modal modules model interactions among modalities. Text–audio interaction can connect lexical polarity with emphasis or prosody, while text–vision interaction can associate descriptions with visible entities, scenes, or facial expressions. Attention weights should be constrained or regularized where possible because a strong correlation does not necessarily imply a causal sentiment cue. Causality aware fusion research highlights the risk of learning spurious cross modal associations (Huang et al., 2025).

At Level 4, the pair wise representations are combined into a global multimodal state. A gated fusion mechanism can estimate context-dependent modality weights. Rather than assigning a fixed contribution to every modality,

The gate can reduce the influence of unreliable or missing channels. This is consistent with recent adaptive fusion and missing-modality research (Liu et al., 2024; Zhao et al., 2025).

At Level 5, the multimodal state is interpreted by an LLM or multimodal foundation model. The LLM can generate a semantic summary, identify potential modality conflicts, and produce an explanation candidate. Importantly, the generated reasoning should be treated as an auxiliary representation that is checked against the underlying evidence. The final prediction can combine discriminative logics with verified semantic features, while explanation generation remains a separate audited output.

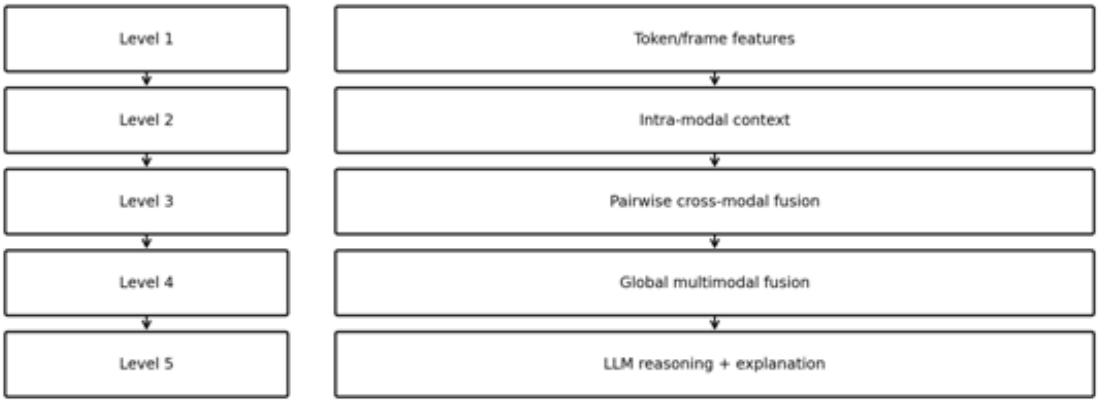

Figure2. Multilevel fusion hierarchy from local evidence to semantic reasoning

5. Generative AI Integration

Generative AI can support the framework at four distinct points: data enrichment, representation enrichment, reasoning support, and explanation. For data enrichment, controlled generation can create paraphrases, image descriptions, or modality-completion candidates. The objective is to expand coverage of linguistic and contextual variation rather than to maximize the number of synthetic records. Synthetic samples should be filtered for semantic consistency and class validity.

For representation enrichment, an LLM can convert non-textual evidence into a textual semantic proxy, such as a sentiment-aware visual caption, while the original modality representation remains available. This dual representation reduces the risk that the generated description becomes the sole source of information. Recent work on sentiment-aware semantic captioning illustrates this design direction for multimodal aspect-based sentiment analysis.

For reasoning support, an LLM can compare modality-specific predictions and identify disagreements. For example, positive lexical wording may coexist with negative facial or acoustic cues. Rather than forcing an early decision, the system can expose the conflict to a semantic reasoning layer and then use uncertainty-aware fusion. This approach makes the LLM a coordinator of evidence rather than a replacement for modality encoders.

For explanation, the system can generate a concise rationale linked to the evidence used by the predictive model. Explanation quality should be evaluated separately from predictive accuracy because a fluent explanation can still be unsupported. A practical audit can compare generated rationales against modality attribution maps, attention distributions, and counterfactual perturbations.

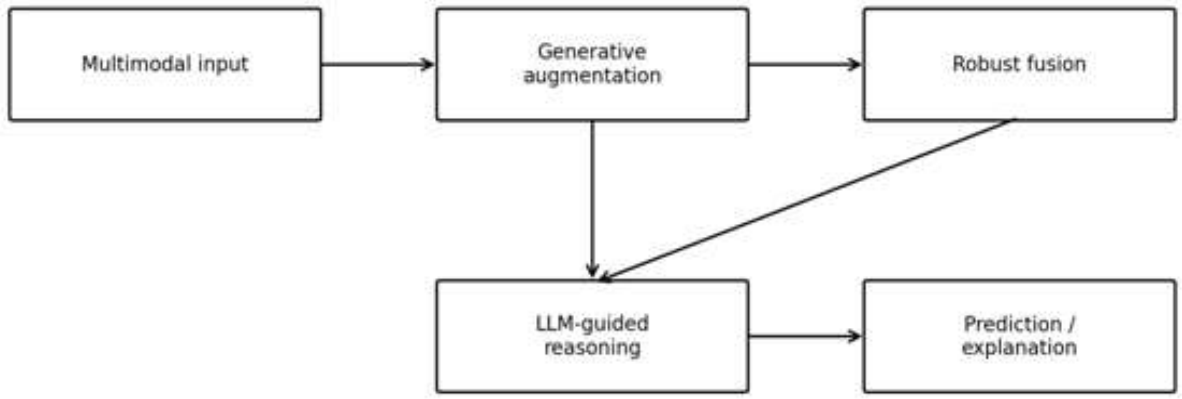

Figure2. Research workflow for GenAI-enhanced multilevel sentiment analysis

6. Methods and Evaluation Framework

The paper adopts a design-oriented synthesis method. Recent peer-reviewed journal literature was examined to identify recurring problems and architectural responses in multimodal sentiment analysis, with emphasis on fusion, hierarchical representation, missing modalities, adaptive weighting, and LLM integration. The resulting framework is a conceptual synthesis rather than a claim of experimental validation.

- **Public Datasets**

Public datasets constitute the empirical foundation of multimodal sentiment analysis research. They enable researchers to compare architectures under relatively standardized experimental conditions and facilitate reproducibility. A public MSA dataset generally contains two or more synchronized modalities, such as text-video, text-audio, text-image, or text-audio-video, together with sentiment or emotion annotations.

Several benchmark datasets have become important for contemporary MSA research. CMU-MOSI contains opinion-oriented YouTube video segments with textual, visual, and acoustic information and continuous sentiment annotations. CMU-MOSEI extends this setting to a much larger collection of video segments involving a broad range of speakers and topics. Recent studies continue to use these datasets as major benchmarks for evaluating multimodal representation learning and fusion mechanisms.

Table I. ***Major Public Datasets Relevant to Multimodal Sentiment Analysis***

Dataset	Main Modalities	Language/Context	Main Annotation	Research Relevance

CMU-MOSI	Text + Audio + Video	English, YouTube opinions	Continuous sentiment	Benchmark for multimodal sentiment and fusion
CMU-MOSEI	Text + Audio + Video	English, YouTube	Sentiment and emotion	Large-scale multimodal benchmark
MELD	Text + Audio + Video	English, conversational	Emotion and sentiment-related labels	Conversational multimodal analysis
CH-SIMS	Text + Audio + Video	Chinese	Continuous sentiment	Chinese multimodal sentiment research
MSAT	Text + Audio + Video	Tamil	Sentiment and emotion	Low-resource language research
DravidianMultiModality	Text + Audio + Visual	Tamil/Malayalam and related contexts	Sentiment/emotion	Regional-language multimodal research
BMSA	Text + Image	Bengali	Positive/negative/neutral	Low-resource image-text sentiment
BanglaMUSE	Text + Audio	Bangla	Positive/negative	Low-resource speech-text sentiment

A future empirical implementation can be evaluated on benchmark datasets such as CMU-MOSI and CMU-MOSEI, supplemented by domain-specific datasets to assess transferability. Evaluation should report accuracy, macro-F1, precision, recall, mean absolute error where regression labels are used, and correlation for continuous sentiment tasks. Robustness experiments should deliberately remove or corrupt one modality and measure performance degradation.

The evaluation should also include calibration and explanation metrics. Expected calibration error can indicate whether confidence values are meaningful. Modality-ablation experiments can reveal whether the model genuinely benefits from multimodal evidence. Synthetic-data audits should compare generated samples with real-data distributions and perform label-consistency checks. Computational measures such as parameter count, latency, memory use, and generation cost are necessary when deployment is intended for real-time systems.

Table II. Recommended evaluation dimensions for multilevel GenAI sentiment systems

Dimension	Measure	Purpose	Interpretation
Classification	Accuracy, macro-F1, precision, recall	Overall predictive quality	Higher values indicate stronger classification under the selected test protocol

Regression	MAE, correlation	Continuous sentiment estimation	Lower MAE and higher correlation generally indicate better agreement
Robustness	Performance under missing/noisy modalities	Tests real-world incompleteness	Smaller degradation indicates greater resilience
Calibration	Expected calibration error	Assesses confidence reliability	Lower error indicates better probability calibration
Generation	Factual/semantic consistency	Audits generated captions and rationales	Higher consistency reduces unsupported reasoning
Efficiency	Latency, memory, compute cost	Deployment feasibility	Lower resource demand improves practical scalability

7. Comparative Analysis of Fusion Strategies

Single-stage early fusion is computationally simple but can obscure modality-specific structure when heterogeneous features are concatenated without sufficient normalization. Late fusion is modular and resilient to some modality differences, yet it can miss fine-grained interactions that occur before independent predictions are produced. Hybrid fusion offers a compromise by combining intermediate interactions with decision-level integration.

Hierarchical and multilevel fusion extends this idea by making the depth of integration explicit. Its central advantage is representational flexibility: shallow features can capture local evidence while deeper layers capture semantic relations. The cost is greater architectural complexity and a larger optimization space. Adaptive gating can partially address this complexity by learning when a modality should contribute strongly or weakly.

LLM-guided fusion adds another abstraction layer. It can provide contextual semantics and natural-language explanations, but it also introduces generation-specific failure modes. Therefore, the proposed framework positions LLM reasoning after evidence-preserving fusion and uses it as an auxiliary semantic mechanism. This ordering helps maintain traceability between the final decision and the original modalities.

8. Challenges and Responsible AI Considerations

Hallucination is a central concern when generated semantic descriptions are used in sentiment analysis. A generated caption may introduce an object, emotion, or event that is not present in the input. Such an error can propagate through the fusion network. Provenance tags, confidence estimates, retrieval constraints, and agreement checks between generated and observed representations can reduce this risk.

Data imbalance and cultural variation also remain important. Sentiment expressions are influenced by language, community norms, sarcasm, humor, and domain-specific conventions. A model trained primarily on one language or platform can therefore exhibit performance differences after deployment. Evaluation should include domain and language splits, subgroup analysis where ethically appropriate, and explicit reporting of data limitations.

Privacy is particularly relevant when audio and visual signals are collected from people. Systems should minimize unnecessary retention, protect sensitive embeddings, obtain appropriate consent, and separate model-development data from personally identifiable information. In customer-feedback applications, the objective should be to understand aggregate experience patterns rather than infer sensitive personal attributes.

Finally, computational sustainability should be considered. Large multimodal models can require substantial memory and energy, while repeated generation can increase operational cost. Distillation, quantization, selective generation, caching, and smaller

task-specific models can be investigated when they preserve acceptable performance and reliability.

9. Discussion

The multilevel perspective reframes fusion as a sequence of decisions about what information should remain separate, what information should interact, and what information should be summarized. This is particularly relevant in the GenAI era because generative models can create a semantic bridge between modalities. However, semantic abstraction should complement—not erase—the underlying evidence.

The literature also indicates a shift from simply adding modalities toward modeling modality relationships. Cross-modal conflicts, missing channels, and modality-specific noise are now explicit design concerns. The proposed framework incorporates these concerns through adaptive weighting, uncertainty-aware reasoning, and separation of observed versus generated evidence. This provides a pathway for building systems that can be evaluated not only by prediction accuracy but also by robustness and traceability.

A second implication is that LLMs should be evaluated as components within a larger architecture. A fluent explanation is not evidence that the prediction is correct. Consequently, future experiments should compare models with and without LLM semantic guidance, quantify the contribution of generated representations, and perform ablations on each fusion level. Such studies would establish whether the additional complexity yields measurable benefits across datasets and domains.

10. Applications

• Customer feedback: combining review text, product images, ratings, emojis, and voice comments can support richer service-quality analysis.

• Social media: multimodal fusion can help distinguish literal sentiment from visually or acoustically contradictory cues.

• Healthcare communication: multimodal sentiment or affective-state analysis can support communication analytics when privacy, consent, and clinical boundaries are explicitly addressed.

• Education and human–computer interaction: voice, facial, and textual signals can provide complementary indicators of engagement or frustration.

• Recommendation systems: sentiment representations can enrich preference modeling when users communicate opinions through multiple media.

11. Conclusion

Generative AI is changing sentiment analysis from a primarily discriminative classification task into a broader problem of multimodal representation, semantic reasoning, and evidence synthesis. The proposed multilevel fusion framework organizes this transition into five stages: local feature preservation, contextual representation, cross-modal interaction, global adaptive fusion, and LLM-guided semantic reasoning. This organization preserves the strengths of conventional multimodal encoders while providing a controlled role for generative models.

The framework also emphasizes that future progress should not be measured only by higher benchmark scores. Robustness to missing modalities, calibration, explanation faithfulness, computational efficiency, privacy, and cross-domain generalization are equally important. Controlled generation and explicit separation of observed from generated evidence can help reduce the risks associated with hallucinated or inconsistent synthetic information.

Overall, multilevel fusion offers a practical conceptual foundation for GenAI-era sentiment systems. Its empirical validation remains an important next step, particularly through controlled ablations, missing-modality experiments, cross-domain testing, and systematic evaluation of generated explanations.

12. Discussion

• Develop uncertainty-aware gates that dynamically distinguish reliable observed evidence from reconstructed or generated evidence.

• Investigate multilingual and code-mixed multilevel fusion, especially for low-resource sentiment settings.

• Design parameter-efficient LLM integration so that semantic reasoning can be deployed under constrained compute budgets.

• Develop explanation audits that measure agreement between generated rationales and modality-level evidence.

• Study continual adaptation under domain shift while preserving privacy and avoiding uncontrolled synthetic-data accumulation.

• Create standardized evaluation protocols that jointly report prediction, robustness, calibration, generation reliability, latency, and energy consumption.

References

[1] Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443. https://doi.org/10.1109/ TPAMI.2018.2798607

[2] Dhotay, M., Dharrao, M., Deokate, S., Bongale, A., & Dharrao, D. (2026). Multimodal sentiment analysis: Emerging innovations, core challenges, and future directions. Discover Artificial Intelligence, 6, 433. https://doi.org/10.1007/s44163-026-01141-2.

[3] Do, H. N., Phan, H. T., & Nguyen, N. T. (2024). Multimodal sentiment analysis using deep learning and fuzzy logic: A comprehensive survey. *Applied Soft Computing, Art. no. 112279*, Article 112279. https://doi.org/10.1016/j.asoc.2024.112279

[4] Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A. (2023). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Information Fusion, 91, 424–444. https://doi.org/10.1016/j.inffus.2022.09.025.

[5] Hou, J., Omar, N., Tiun, S., Saad, S., & He, Q. (2024). TCHFN: Multimodal sentiment analysis based on text-centric Hierarchical Fusion Network. *Knowledge-Based Systems*, *300*, Article 112220. https://doi.org/10.1016/j.knosys.2024.112220

[6] Huang, C., Chen, J., Huang, Q., Wang, S., Tu, Y., & Huang, X. (2025). AtCAF: Attention-based causality-aware fusion network for multimodal sentiment analysis. Information Fusion, 114, 102725. https://doi.org/10.1016/j.inffus.2024.102725

[7] Huang, C., Zhang, J., Wu, X., Wang, Y., Li, M., & Huang, X. (2023). TeFNA: Text-centered fusion network with crossmodal attention for multimodal sentiment analysis. Knowledge-Based Systems, 269, 110502. https://doi.org/10.1016/j.knosys.2023.110502

[8] Kim, Y., & Bartolini, I. (2026). Reframing sentiment analysis in the LLM era: A comprehensive review of methods, evaluation, and sustainability. Artificial Intelligence Review. https://doi.org/10.1007/s10462-026-11684-6

[9] Krugmann, J. O., & Hartmann, J. (2024). Sentiment analysis in the age of Generative AI. *Customer Needs and Solutions*, *11*, Article 3. https://doi.org/10.1007/s40547-024-00143-4

[10] Lai, S., Hu, X., Xu, H., Ren, Z., & Liu, Z. (2023). Multimodal sentiment analysis: A survey. *Displays*, *80*, Article 102563. https://doi.org/10.1016/j.displa.2023.102563

[11] Li, Y., Ding, H., Lin, Y., Feng, X., & Chang, L. (2024). Multi-level textual-visual alignment and fusion network for multimodal aspect-based sentiment analysis. Artificial Intelligence Review, 57, 78. https://doi.org/10.1007/s10462-023-10685-z

[12] Li, Z., Huang, Z., Pan, Y., Yu, J., Liu, W., Chen, H., Luo, Y., & Wu, D. (2024). Hierarchical denoising representation disentanglement and dual-channel cross-modal-context interaction for multimodal sentiment analysis. Expert Systems with Applications, 252, 124236. https://doi.org/10.1016/j.eswa.2024.124236

[13] Li, S., & Tang, H. (2026). Multimodal alignment and fusion: A survey. *International Journal of Computer Vision*, *134*, Article 103. https://doi.org/10.1007/s11263-025-02667-1

[14] Liu, Z., Cai, L., Yang, W., & Liu, J. (2024). Sentiment analysis based on text information enhancement and multimodal feature fusion. Pattern Recognition, 156, 110847. https://doi.org/10.1016/j.patcog.2024.110847

[15] Liu, Z., Zhou, B., Chu, D., Sun, Y., & Meng, L. (2024). Modality translation-based multimodal sentiment analysis under uncertain missing modalities. Information Fusion, 101, 101973. https://doi.org/10.1016/j.inffus.2023.101973

[16] Mu, J., Wang, W., Liu, W., Yan, T., & Wang, G. (2025). Multimodal large language model with LoRA fine-tuning for multimodal sentiment analysis. *ACM Transactions on Intelligent Systems and Technology*, *16*(6), 139:1–139:23. https://doi.org/10.1145/3709147

[17] Pandey, A., & Vishwakarma, D. K. (2023). Progress, achievements, and challenges in multimodal sentiment analysis using deep learning: A survey. *Applied Soft Computing*, *152*, Article 111206. https://doi.org/10.1016/j.asoc.2023.111206.

[18] Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing research

based on function-component-content relations. Information Fusion, 37, 98–125. https://doi.org/10.1016/j.inffus.2017.02.003

[19] Rupapara, V., Rustam, F., Shahzad, H. F., Mehmood, A., Ashraf, I., & Choi, G. S. (2021). Impact of SMOTE on imbalanced text features for toxic comments classification using RVVC model. *IEEE Access*, *9*, 78621–78634.

[20] Singh, S., Kumar, K., & Kumar, B. (2026). A fusion of social media context for efficient multimodal sentiment analysis using deep learning techniques. Discover Artificial Intelligence, 6, 329. https://doi.org/10.1007/s44163-026-00986-x

[21] Tzirakis, P., Chen, J., Zafeiriou, S., & Schuller, B. (2020). End-to-end multimodal affect recognition in real-world environments. *Information Fusion*, *68*, 46–53. https://doi.org/10.1016/j.inffus.2020.10.011

[22] Wang, B., Dong, G., Zhao, Y., Li, R., Yin, W., & Lu, L. (2025). Tripartite interaction representation learning for multi-modal sentiment analysis. Expert Systems with Applications, 268, 126279. https://doi.org/10.1016/j.eswa.2024.126279

[23] Wang, L., Peng, J., Zheng, C., Zhao, T., & Zhu, L. (2024). A cross modal hierarchical fusion multimodal sentiment analysis method based on multi-task learning. Information Processing & Management, 61(3), 103675. https://doi.org/10.1016/j.ipm.2024.103675

[24] Yang, H., Zhao, Y., Wu, Y., Wang, S., Zheng, T., Zhang, H., Ma, Z., Che, W., Wang, S., Wei, S., & Qin, B. (2025). Large language models meet text-centric multimodal sentiment analysis: A survey. *Science China Information Sciences*, *68*(10), Article 200101. https://doi.org/10.1007/s11432-024-4593-8

[25] Yi, G., Fan, C., Zhu, K., Lv, Z., Liang, S., Wen, Z., Pei, G., Li, T., & Tao, J. (2023). VLP2MSA: Expanding vision-language pre-training to multimodal sentiment analysis. *Knowledge-Based Systems*, *283*, Article 111136. https://doi.org/10.1016/j.knosys.2023.111136

[26] Zhao, X., Poria, S., Li, X., Chen, Y., & Tang, B. (2025). Toward robust multimodal sentiment analysis using multimodal foundational models. Expert Systems with Applications, 276, 126974. https://doi.org/10.1016/j.eswa.2025.126974

[27] Zhu, C., Chen, M., Zhang, S., Sun, C., Liang, H., Liu, Y., & Chen, J. (2023). SKEAFN: Sentiment knowledge enhanced attention fusion network for multimodal sentiment analysis. Information Fusion, 100, 101958. https://doi.org/10.1016/j.inffus.2023.101958

[28] Zhu, L., Zhao, H., Zhu, Z., Zhang, C., & Kong, X. (2025). Multimodal sentiment analysis with unimodal label generation and modality decomposition. Information Fusion, 116, 102787. https://doi.org/10.1016/j.inffus.2024.102787

[29] Zhu, L., Zhu, Z., Zhang, C., Xu, Y., & Kong, X. (2023). Multimodal sentiment analysis based on fusion methods: A survey. Information Fusion, 95, 306–325. https://doi.org/10.1016/j.inffus.2023.02.028

[30] Zhu, M., et al. (2026). Multi-level global and local fusion for multimodal sentiment analysis. Memetic Computing. https://doi.org/10.1007/s12293-026-00520-7