STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–15.
Published online in International Academic Press (www.IAPress.org)

Multimodal Arabic Sentiment Analysis: A Comprehensive Review

Ayoub Ben Cheikhi*, ELl Habib Nfaoui

L3IA Laboratory, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez, 30000, Morocco

Abstract Multimodal Sentiment Analysis integrates textual, visual, and auditory information to achieve richer contextual understanding. Although multimodal learning has been advanced significantly in high-resource languages, research in Arabic remains limited and fragmented. This paper presents a comprehensive review of Multimodal Arabic Sentiment Analysis, to examine datasets, modeling paradigms, and fusion strategies. We survey available multimodal resources, analyze transformer-based architectures and large pretrained models across language, vision, and audio, and compare fusion mechanisms. Our review highlights persistent challenges, including limited dataset scale, dialectal diversity, modality imbalance favoring text, and trade-offs between modeling expressiveness and computational efficiency. Critically, existing fusion mechanisms consistently fail to robustly handle dialectal variance, as most architectures implicitly assume standardized linguistic representations and therefore struggle to generalize across the rich spectrum of Arabic dialects. We conclude by outlining key research gaps and future directions toward scalable, dialect-aware, and interaction-centric multimodal systems for the Arabic linguistic ecosystem.

Keywords Multimodal Sentiment Analysis, Transformer Architectures,Large Pretrained Models, Multimodal Datasets, Dialectal Arabic

DOI: 10.19139/soic-2310-5070-3641

1. Introduction

Multimodal Arabic Sentiment Analysis has recently emerged as a transformative paradigm that enables computational systems to process, integrate, and interpret information derived from multiple data modalities, such as textual content, speech signals, and visual inputs [1]. Unlike unimodal approaches that rely on a single source of data [2, 3], multimodal frameworks leverage the complementary nature of diverse modalities to achieve a richer and more holistic contextual understanding. The rapid advancement of transformer-based architectures, coupled with the development of large-scale pretrained models — including Large Language Models (LLMs) [4, 5], Large Vision Models (LVMs) [6], and Large Audio Models (LAMs) [7] — has significantly accelerated progress in multimodal learning and representation.

Within this evolving landscape, the Arabic language remains comparatively underexplored, particularly in Multimodal Arabic Sentiment Analysis settings. While significant progress has been achieved in multimodal learning for high-resource languages [8], research efforts addressing Arabic across integrated textual, visual, and auditory modalities remain limited and fragmented. This paper therefore presents a comprehensive review of multimodal Arabic AI applications that leverage transformer architectures alongside large pretrained models spanning language, vision, and audio domains. To provide a structured analytical lens, sentiment analysis is adopted as a representative case study through which existing methodologies, datasets, and evaluation strategies are examined and compared.

Sentiment analysis [9], also known as opinion mining, focuses on the automatic identification and classification of emotions, attitudes, and subjective opinions expressed in human communication. It plays a critical role in

*Correspondence to: Ayoub Ben Cheikhi (Email: ayoub.bencheikhi@usmba.ac.ma). L3IA Laboratory, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez, 30000, Morocco

ISSN 2310-5070 (online) ISSN 2311-004X (print)

numerous real-world applications, including social media monitoring, customer feedback analysis, and decision-support systems. However, conducting sentiment analysis in Arabic presents unique challenges. The language is characterized by rich morphology, complex syntactic structures, extensive dialectal variation, and a scarcity of large, annotated multimodal datasets. These factors collectively complicate model training, generalization, and evaluation, thereby motivating the need for advanced multimodal approaches tailored specifically to the Arabic linguistic context.

The remainder of this paper is organized as follows. Section 2 presents the theoretical background of multimodal AI and large pretrained models. Section 3 reviews Arabic multimodal datasets. Section 4 discusses fusion strategies and semantic alignment. Section 5 examines the evolution of multimodal sentiment analysis in Arabic, while Section 6 provides a critical analysis of existing approaches. Section 7 analyzes fusion mechanisms under Arabic-specific constraints. Section 8 highlights key research gaps, and Section 9 outlines future research directions. Finally, Section 10 concludes the paper.

2. Background and Theoretical Foundations

This section reviews the evolution of transformer architectures, and examines the rise of large pretrained models. It further discusses the complementary roles of Large Language Models (LLMs), Large Vision Models (LVMs), and Large Audio Models (LAMs) within multimodal learning ecosystems, establishing the conceptual basis for advanced cross-modal intelligent systems.

2.1. Transformer Architectures

Transformer architectures have become a foundational paradigm in modern artificial intelligence, driving substantial progress across natural language processing, computer vision, and generative modeling. Originating from sequence transduction tasks, transformers rapidly evolved due to their self-attention mechanisms, which enable efficient modeling of long-range dependencies and parallelizable training. [10] provide a comprehensive historical survey tracing the development of transformer-based models from their early language-focused designs to their widespread adoption in diverse domains, including text generation, vision applications, and generative deep learning. Their analysis combines qualitative and quantitative perspectives, highlighting key architectural innovations, scaling strategies, and optimization techniques that have shaped the transformer ecosystem. The authors further emphasize emerging research directions, particularly the expansion toward multimodal learning and improvements in training efficiency.

2.2. Evolution of Large Pretrained Models

The evolution of large pretrained models marks a pivotal shift in artificial intelligence, characterized by the transition from task-specific training toward highly generalized, large-scale representation learning. These models, pretrained on massive and diverse datasets, have demonstrated the capacity to acquire transferable knowledge that can be adapted to a wide spectrum of downstream applications. Lehman et al. [11] explore this paradigm through the lens of evolutionary machine learning, illustrating how large language models trained for code generation can augment genetic programming processes. Their work shows that such models are capable of guiding mutation operators via human-like program transformations, thereby improving the efficiency and creativity of program evolution. By leveraging learned priors embedded within large models, evolutionary systems can generate extensive sets of functional program variants and even bootstrap the development of new models. This integration reveals broader implications for open-ended learning frameworks, as well as for the advancement of deep learning and reinforcement learning methodologies. Collectively, these insights highlight how the convergence of large pretrained models and evolutionary strategies is fostering more adaptive, scalable, and autonomous intelligent systems.

2.3. The Roles of Large Language Models (LLMs), Large Vision Models (LVMs), and Large Audio Models (LAMs) in Multimodal Learning Ecosystems

Large Language Models (LLMs) provide advanced linguistic reasoning, contextual understanding, and generative capabilities that enable semantic alignment across modalities. Extending these capabilities, multimodal large language models integrate visual inputs alongside textual representations, facilitating complex cross-modal reasoning tasks. [12] survey the rapid evolution of such Multimodal Large Language Models (MLLMs), highlighting their emergent abilities to unify vision–language processing beyond traditional multimodal pipelines. Their review examines architectural designs, large-scale training strategies, and evaluation frameworks, while also identifying persistent challenges, including hallucination, reliability, and multimodal chain-of-thought reasoning.

Complementing LLMs, Large Vision Models (LVMs), particularly Vision–Language Models (VLMs), enable deep visual perception integrated with semantic interpretation. [13] provide a comprehensive survey of state-of-the-art VLMs, emphasizing their superiority over unimodal vision systems in tasks requiring joint visual–textual reasoning. They analyze alignment mechanisms, benchmarking protocols, and architectural innovations, while also discussing critical concerns such as fairness, safety, hallucination, and cross-modal consistency.

In the auditory domain, Large Audio Models (LAMs) expand multimodal ecosystems by modeling speech signals, paralinguistic patterns, and affective vocal cues. [14] investigate the application of speech and language foundation models for automatic depression detection, integrating transcription systems with speech and text embeddings optimized through low-rank adaptation techniques. Their findings demonstrate that multimodal audio–language foundation models outperform traditional baselines across multilingual datasets, even under moderate transcription quality constraints.

Collectively, the integration of LLMs, LVMs, and LAMs enables holistic multimodal representation learning, fostering systems capable of perceiving, reasoning, and interacting across language, vision, and audio domains. This convergence underpins the development of more context-aware, human-centric, and cognitively aligned artificial intelligence systems.

3. Arabic Multimodal Resources and Datasets

The development of multimodal artificial intelligence systems relies fundamentally on the availability of large-scale, well-annotated datasets that integrate multiple information streams. In the Arabic context, however, multimodal resources remain comparatively limited in scale, diversity, and accessibility. This section surveys existing Arabic multimodal datasets and benchmarks, analyzing them in terms of modality composition, application domains, annotation practices, and public availability. Table 1 summarizes the most prominent resources discussed in this review.

Early efforts in Arabic multimodal resource construction focused primarily on document and speech integration. The Multi-Modal Arabic Corpus (MMAC) introduced by [15] combined textual data with annotated document images to support linguistic analysis and optical character recognition (OCR) development. Similarly, the AVAS corpus proposed by. [16] provided an audio-visual speech database designed for multimodal recognition applications such as lip-reading and audio-visual speech recognition under challenging recording conditions.

Subsequent work expanded toward affective computing and social media analysis. [17] developed AVANEmo, an audio-visual dataset for Arabic emotion recognition comprising annotated video and audio clips covering six basic emotions. [18] further advanced this direction by introducing a natural-emotion audio-visual dataset and demonstrating high multimodal recognition performance using deep learning models. In parallel, multimodal sentiment and misinformation analysis emerged as key research areas. [19] proposed an Arabic multimodal sentiment dataset validated using SVM classifiers, while [20] investigated rumor detection through joint textual and visual tweet representations.

More recent datasets emphasize linguistic diversity and dialectal representation. For example, [21] introduced EGY-MER, the first multimodal emotion recognition dataset for Egyptian Arabic, integrating speech, text, and facial cues. [22] proposed MDER-MA for Moroccan Arabic emotion recognition, addressing low-resource dialect challenges through multimodal annotation. While these datasets represent an important step toward incorporating

dialectal diversity, they remain limited to single-dialect settings (e.g., Egyptian Arabic in EGY-MER and Moroccan Arabic in MDER-MA). Notably, existing works do not investigate cross-dialect generalization, where models trained on one dialect are evaluated on another. This limitation restricts our understanding of model robustness in real-world multilingual and multi-dialect Arabic scenarios.

Table 1. Summary of Arabic Multimodal Datasets. Annotation quality is indicated when Inter-Annotator Agreement (IAA) is explicitly reported; otherwise, it is marked as Not Reported, highlighting transparency limitations in current resources.

Dataset	**Mod.**	**Applications / Features**	**Availability**	**Annotation (IAA)**
MMAC (2010) [15]	T, I	OCR, linguistic research	Private	Not Reported
AVAS (2013) [16]	A, V	AVSR, lip-reading	Private	Not Reported
AVANEmo (2019) [17]	A, V	Emotion recognition (6 emotions)	Not available	Not Reported
Bellagha & Zrigui (2021) [23]	A, T	Speaker role recognition	Public	Not Reported
Alqarafi et al. (2017) [19]	T, V	Sentiment analysis	Not available	Not Reported
Albalawi et al. (2023) [20]	T, SP	Rumor detection	Not available	Not Reported
ArabSign (2023) [24]	V, D, S	ArSL recognition	Public	Manual Annotation
Al Roken et al. (2023) [18]	A, V	Emotion recognition	Private	Not Reported
Abbas et al. (2024) [25]	A, T, V	ArSL, religious speech	Not available	Not Reported
L2-KSU (2025) [26]	A, T	Mispronunciation detection	Not available	Not Reported
EGY-MER (2025) [21]	A, T, V	Emotion recognition (Egyptian Arabic)	Not available	Human Annotators
MDER-MA (2025) [22]	A, T, V	Emotion recognition (Moroccan Arabic)	Public	Human Annotators
Ar-MuSA (2025) [27]	A, T, V	Sentiment analysis	Public	IAA Reported
MSAD–Darija (2026) [28]	A, T, V	Dialect sentiment analysis	Public	Native Annotators
CAMEL-Bench (2025) [29]	T, I	Multimodal LMM benchmarking	Public	Not Reported

Mod. = Modalities; A = Audio; T = Text; V = Video; I = Image; D = Depth; S = Skeleton joints; SP = Social Media Post. "Not Reported" indicates that no explicit Inter-Annotator Agreement (IAA) score was provided in the original publication.

Complementing this effort, Our work [28] introduces MSAD–Moroccan Dialect, the first multimodal sentiment analysis dataset for Darija, built from podcast videos integrating aligned audio, text, and visual features. Annotated by native speakers and benchmarked using SVM, the dataset provides a foundational resource for advancing sentiment modeling in low-resource Arabic dialects. Sign language and accessibility resources also constitute an important research direction. [25] introduced a multimodal Arabic Sign Language dataset integrating sign videos, speech, and textual annotations within religious discourse. Similarly, [24] developed ArabSign, a large-scale continuous sign language dataset incorporating color, depth, and skeletal modalities for recognition benchmarking. Additional domain-specific datasets support speech learning and broadcast analysis. [26] constructed a speech dataset for transformer-based mispronunciation detection in computer-assisted language learning, while [23] introduced a multimodal corpus for speaker role recognition in Arabic television broadcasts, integrating audio and textual streams. Benchmarking initiatives have also begun to emerge. CAMEL-Bench, proposed by [29], provides a comprehensive evaluation framework for Arabic multimodal large models across visual reasoning and understanding tasks. In sentiment analysis, [27] presented Ar-MuSA, an open multimodal dataset combining text, audio, and visual modalities, demonstrating improved classification performance through multimodal fusion. [30] likewise explored transformer-based pipelines for constructing Arabic multimodal sentiment datasets despite limited data availability.

Overall, the surveyed resources demonstrate growing momentum in Arabic multimodal dataset development, spanning applications in affective computing, accessibility, speech learning, misinformation detection, and multimodal benchmarking. Nevertheless, A critical examination of the surveyed datasets reveals that most multimodal models are effectively dialect-locked, as they are both trained and tested on data from the same dialect. To date, there is no systematic evaluation of cross-dialect transfer in Arabic multimodal sentiment or

emotion recognition. Consequently, it remains unclear how well these models generalize across dialectal variations, particularly given the significant lexical, phonological, and semantic differences among Arabic dialects.

4. Semantic Alignment in Multimodal Learning

Multimodal learning is not merely a question of architectural categorization but fundamentally a problem of semantic alignment across heterogeneous representational spaces. While the literature traditionally classifies fusion mechanisms into early, late, hybrid, tensor-based, and attention-driven strategies [31, 32, 33, 34, 35, 36, 37], such taxonomies often obscure a deeper issue: whether modalities are semantically grounded or simply aggregated. Table 2 summarizes the surveyed fusion strategies and their primary contributions.

4.1. Beyond Concatenation: From Multi-Stream to True Multimodal Learning

Early fusion methods commonly integrate modality-specific embeddings prior to classification [31, 32]. Although contrastive objectives [31] and modality translation mechanisms [32] enhance representational robustness, many implementations remain dependent on embedding concatenation. However, concatenating embeddings extracted from models such as MARBERT and Vision Transformers does not inherently establish semantic correspondence between linguistic and visual concepts. Rather, this strategy constitutes *multi-stream learning*, where modalities coexist in a shared vector space without explicit conceptual interaction. Similarly, tensor-based fusion approaches, including the Tensor Fusion Network [34], HyCon [35], and TensorFormer [36], model higher-order interactions through structured representations. While these methods capture intra- and inter-modal correlations, they do not always guarantee semantic grounding between culturally specific linguistic constructs and visual expressions. The representational interaction is mathematically defined, yet the conceptual linkage between modalities remains implicitly learned rather than explicitly aligned.

4.2. Limitations of Decision-Level Integration

Late fusion techniques combine modality-specific predictions at the decision level [33, 38]. Architectures such as MMLATCH [33] introduce hierarchical masking strategies to enhance cross-modal supervision, while benchmark initiatives such as MuSe 2023 [38] demonstrate competitive performance using GRU-based ensembles. In Arabic contexts, multimodal sarcasm detection frameworks [1] have shown substantial gains over unimodal baselines. Nevertheless, decision-level aggregation does not resolve the semantic misalignment problem. Modalities are processed independently and reconciled only after inference, limiting the model's ability to construct a unified semantic interpretation. Such approaches may improve predictive accuracy yet fail to ensure that visual gestures, tonal patterns, and linguistic expressions are conceptually coherent within a shared semantic framework.

4.3. Cultural Specificity and Semantic Grounding in Arabic

The challenge of semantic alignment becomes more pronounced in morphologically rich and culturally grounded languages such as Arabic. Figurative expressions, pragmatic devices, and rhetorical constructs—such as *Tawriyah* (double entendre)—often require contextual and cultural knowledge beyond surface-level embeddings. Cross-modal attention architectures [37, 39] represent a significant step toward interaction-aware modeling by enabling modalities to influence each other dynamically. For example, GCMA-Net [39] demonstrates improved cross-modal reasoning in Arabic multimodal sentiment analysis through gated attention mechanisms. However, even attention-based models may struggle to interpret culturally specific gestures or idiomatic expressions without access to structured external knowledge. Neural alignment alone may capture statistical dependencies but cannot always encode sociocultural semantics embedded in gestures, tone, or dialectal nuance. Consequently, purely neural fusion strategies—whether early, tensor-based, contextual [40, 41], or attention-driven—remain limited in their ability to incorporate culturally grounded semantic structures.

4.4. Toward Knowledge-Aware Semantic Alignment

Reframing fusion around semantic alignment shifts the focus from how modalities are computationally merged to how meaning is jointly constructed. Effective multimodal systems should:

- Establish explicit correspondences between linguistic tokens and visual or acoustic cues;
- Regulate cross-modal influence through structured interaction mechanisms;
- Mitigate modality imbalance while preserving conceptual coherence;
- Integrate culturally informed knowledge representations when necessary.

In summary, while prior work has progressively advanced multimodal fusion architectures [31, 32, 33, 34, 35, 36, 37, 39], the core challenge remains semantic grounding rather than architectural complexity. True multimodal learning in Arabic sentiment analysis requires models that explicitly align meaning across modalities and incorporate cultural knowledge structures beyond simple neural aggregation.

Table 2. Summary of Multimodal Fusion Strategies

Category	**Method**	**Key Idea**	**Contribution**
Early Fusion	Nguyen et al. (2023) [31] Liu et al. (2024) [32]	Angular contrastive learning Modality translation	Enhanced embedding alignment Robust to missing modalities
Late Fusion	MMLATCH (2022) [33] MuSe 2023 [38] Our (2025) [1]	Bottom-up/top-down masking GRU baselines BiLSTM + CNN + ResNet	SOTA on CMU-MOSEI benchmark results 97% accuracy (Arabic sarcasm)
Hybrid Fusion Tensor Fusion	CRNN-SVM (2023) [42] TFN (2017) [34]	Feature + decision integration High-order tensor interaction	91% average accuracy Captures inter-modal dynamics
Tensor + Contrastive	HyCon (2022) [35] TensorFormer (2022) [36]	Contrastive tri-modal learning Tensor-based transformer	Reduced modality gap Simultaneous modality interaction
Contextual Fusion	TATE (2022) [40] SUGRM (2023) [41]	Tag-assisted alignment Self-supervised recalibration	Missing modality handling Improved representation learning
Cross-Modal Attention	SKESL (2023) [37] GCMA-Net (2026) [39]	Knowledge-enhanced SSL Gated cross-modal attention	Mitigates overfitting SOTA in Arabic MSA

5. Evolution of Multimodal Sentiment Analysis in Arabic

The evolution of multimodal sentiment analysis in Arabic contexts reflects a progressive shift from handcrafted feature-based approaches toward interaction-aware deep learning architectures. This transition is driven not only by advances in modeling techniques but also by the increasing availability of multimodal datasets and pretrained representations.

Early work by [43] established one of the first multimodal baselines for Arabic video sentiment analysis. Their framework combined linguistic, acoustic, and facial features using traditional machine learning classifiers, achieving 76% accuracy. However, the approach relied heavily on handcrafted descriptors and shallow models, without leveraging contextualized language representations or temporal visual dynamics. Fusion was limited to feature concatenation, which does not explicitly model cross-modal dependencies and is therefore insufficient for capturing complex interactions between modalities.

A more systematic exploration of fusion strategies was introduced by [44], who evaluated feature-level, score-level, and decision-level fusion within a unified framework (Figure 1). Their hybrid feature–score fusion achieved 94.08% accuracy on the SADAM dataset. Despite this strong performance, the small dataset size (524 utterances) raises concerns about generalizability and potential overfitting. Moreover, the reliance on static Word2Vec embeddings for text limits the model's ability to capture contextual semantics, particularly in morphologically rich and dialectally diverse Arabic.

The introduction of the Ar-MuSA dataset by [27] represents a major milestone in benchmarking Arabic multimodal sentiment analysis. Unlike earlier datasets, Ar-MuSA provides a larger and more diverse multimodal

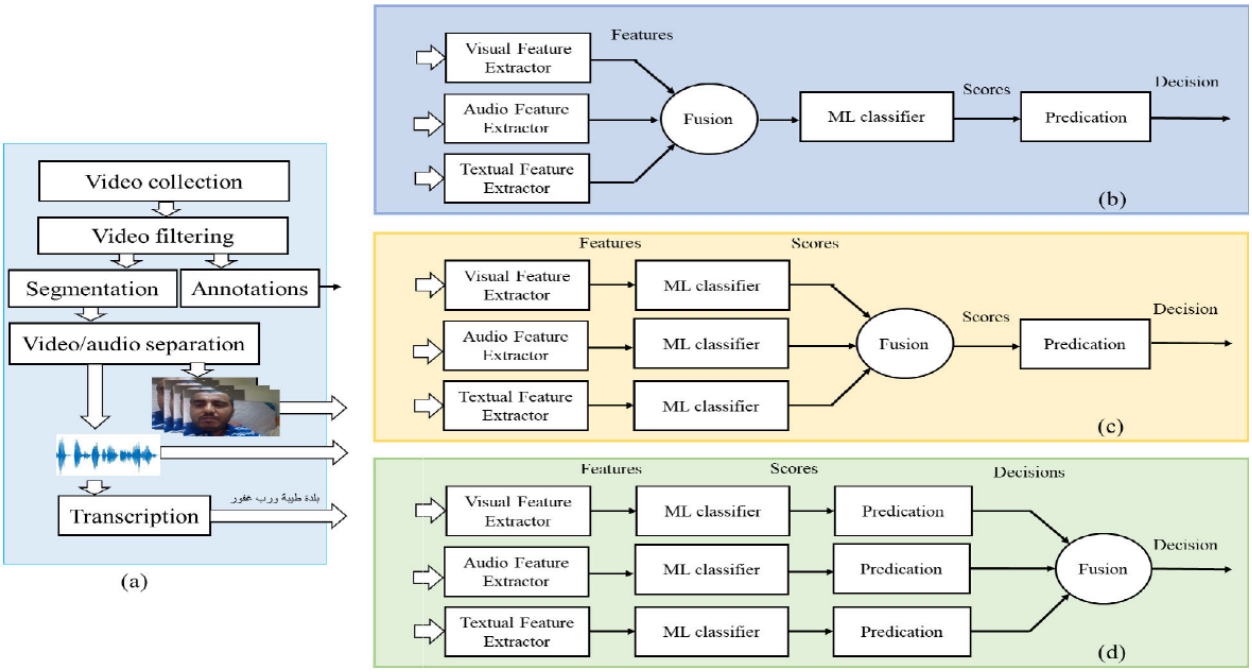

Figure 1. Multimodal video analytics framework illustrating feature-level, score-level, and decision-level fusion strategies [44].

corpus, enabling more realistic evaluation. Their results highlight a fundamental challenge: modality imbalance. Text-only models achieve 71% F1-score, significantly outperforming audio (39%) and image (45%). Although late fusion improves audio performance (up to 67% F1), multimodal large language model (LLM) approaches achieve only 60% F1, indicating limited cross-modal synergy and reinforcing the dominance of textual representations.

To address data scarcity and modality misalignment, [45] proposed UniTextFusion, which reformulates multimodal inputs into a unified textual representation (Figure 2). Audio and visual signals are converted into descriptive text and combined with transcripts for LoRA-based fine-tuning of large language models. While this approach is computationally efficient and achieves up to 71% F1-score, its reliance on textualization compresses fine-grained acoustic and visual information and does not explicitly model cross-modal interactions.

Building on transformer-based representations, [46] introduced an early-fusion framework that integrates embeddings from MarBERT, HuBERT, and ViT, followed by SVM classification (Figure 3). This approach achieves 77.6% F1-score on Ar-MuSA, outperforming several baseline methods. However, the use of simple embedding concatenation limits the model's ability to capture higher-order dependencies between modalities, particularly under conditions of modality noise and imbalance.

To overcome these limitations, GCMA-Net [39] introduces a gated cross-modal attention mechanism (Figure 4). The model projects modality-specific embeddings into a shared latent space and applies cross-modal attention to capture interdependencies, followed by reliability-aware gated fusion. This architecture achieves 80.17% accuracy and 0.8015 F1-score on Ar-MuSA, significantly outperforming both early and late fusion approaches. Importantly, ablation studies demonstrate that both cross-modal attention and gating contribute to performance gains, highlighting the importance of interaction-aware modeling.

Despite these advancements, several persistent limitations remain across the literature. First, **dataset scale and generalizability** continue to constrain model performance, as most Arabic multimodal datasets are relatively small compared to their English counterparts. Second, **modality imbalance** persists, with textual representations consistently dominating predictive performance, while audio and visual modalities provide weaker and often noisy signals. Third, **fusion design constraints** limit the effectiveness of multimodal learning: early fusion fails to capture

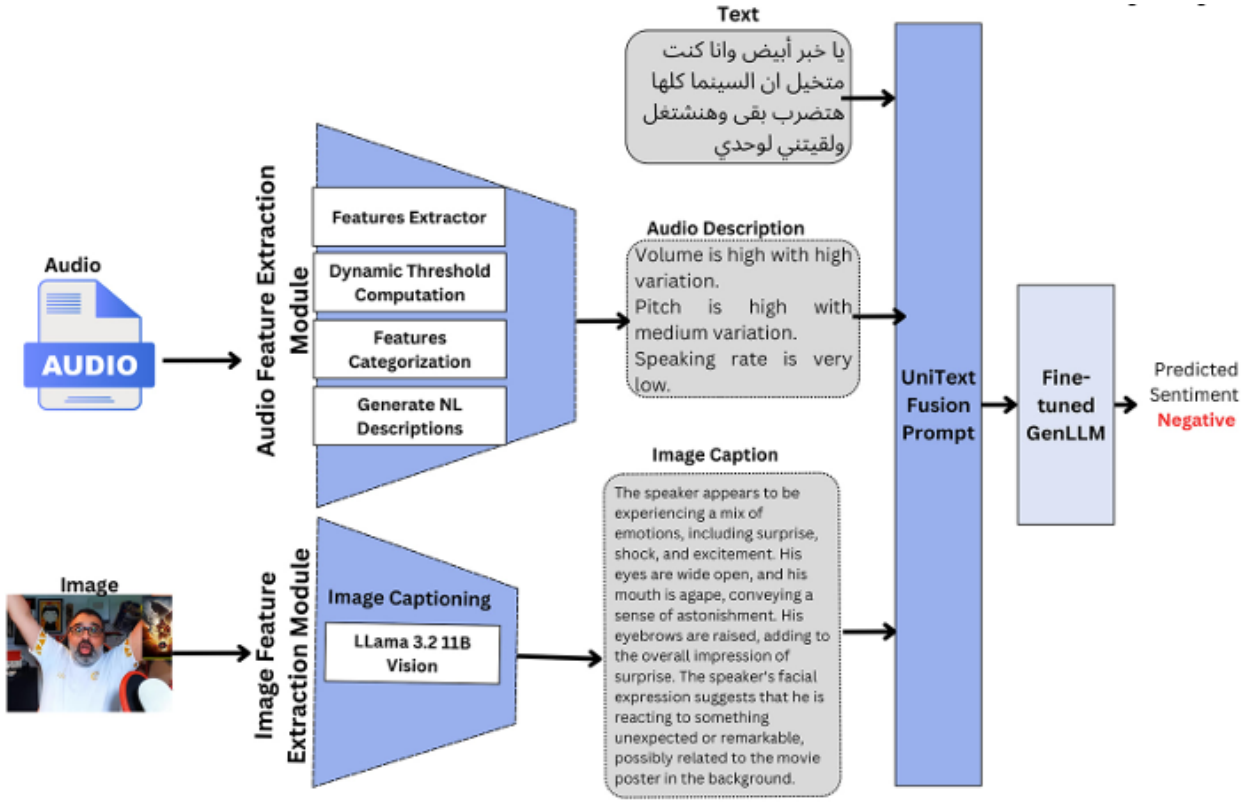

Figure 2. UniTextFusion early textual fusion pipeline converting audio and image signals into descriptive textual prompts [45].

complex dependencies, late fusion cannot resolve modality conflicts, and interaction-aware architectures, while more effective, introduce additional computational complexity.

Overall, the progression from handcrafted feature concatenation [43] to gated cross-modal attention [39] reflects a growing recognition that effective multimodal sentiment analysis in Arabic requires explicit modeling of cross-modal interactions. However, architectural improvements alone are insufficient. Achieving robust, scalable, and modality-balanced systems remains an open challenge, particularly in low-resource and dialectally diverse Arabic environments.

6. Critical Diagnostic

6.1. Methodological Flaws

Despite measurable progress in Arabic multimodal sentiment analysis (MSA), several recurring methodological weaknesses limit the robustness and interpretability of existing systems. A number of studies rely either on handcrafted representations or pretrained architectures without thoroughly addressing Arabic-specific adaptation and dialectal variability; for example, [43] employs handcrafted audio–visual features and traditional classifiers, while [44] adopts conventional fusion strategies without explicitly mitigating potential cross-lingual transfer bias. More recent frameworks such as [27] and [45] integrate pretrained or generative models, yet the extent of linguistic adaptation and resistance to negative transfer remains insufficiently validated. In addition, controlled modality-level ablation is frequently absent: although unimodal and multimodal comparisons are reported in [43] and [46], these evaluations do not rigorously quantify the independent marginal contribution of audio and visual modalities when

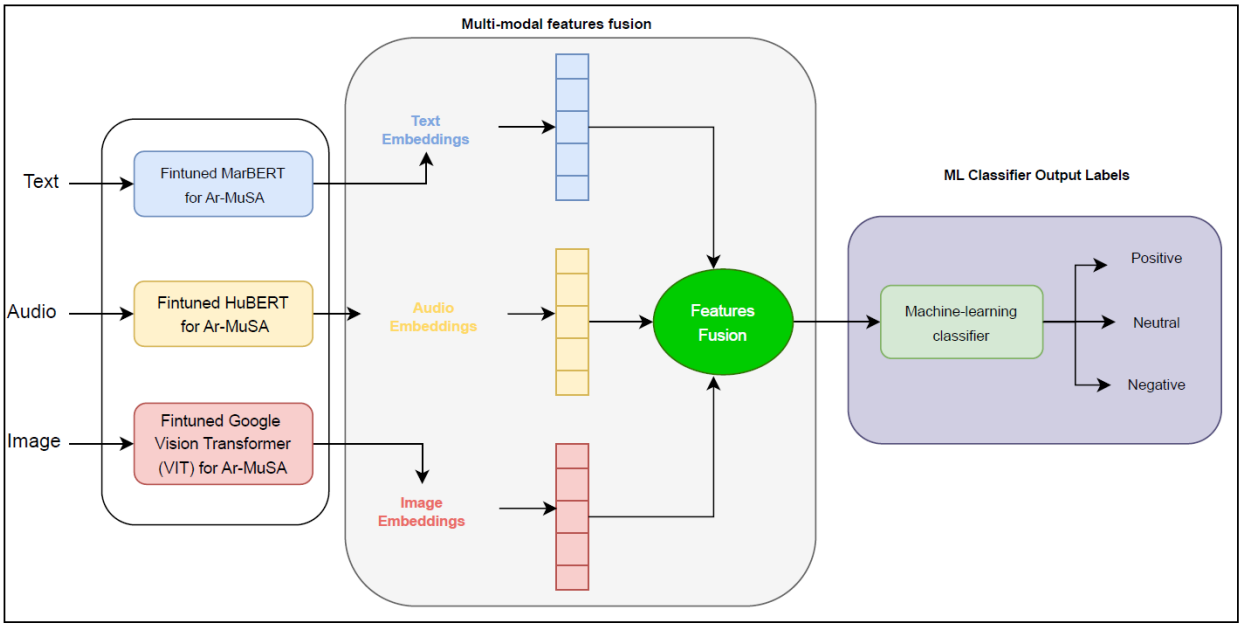

Figure 3. Transformer-based early fusion architecture with modality-specific encoders and SVM classifier [46].

textual input is present, leaving open the possibility that reported gains stem primarily from textual dominance rather than genuine cross-modal complementarity. Furthermore, attention-based fusion architectures are typically treated as intrinsically effective, yet studies such as [46] and [45] do not provide qualitative interpretability analyses (e.g., attention visualization or token-level attribution) to demonstrate that cross-modal attention captures sentiment-bearing cues rather than superficial co-occurrence patterns, thereby reinforcing the "black-box" nature of these systems. Finally, evaluation practices across [43, 44, 27, 45, 46] predominantly emphasize Accuracy and weighted F1-score, which may obscure performance degradation under class imbalance and prediction uncertainty; therefore, future research should incorporate robustness-oriented metrics such as AUC-ROC and calibration error measures to ensure both discriminative strength and probabilistic reliability.

6.2. Analytical Synthesis and Cross-Study Comparison

Rather than viewing existing work as incremental improvements in model design, we argue that Arabic multimodal sentiment analysis is fundamentally constrained by three interacting factors: *(i) linguistic complexity and dialectal variability*, *(ii) modality reliability imbalance*, and *(iii) fusion mechanism limitations*. This section reframes prior studies through this analytical lens and evaluates how different architectures respond to these challenges.

6.2.1. A Taxonomy of Challenges in Arabic Multimodal Sentiment Analysis Based on the surveyed literature, we identify three core challenges that fundamentally constrain the effectiveness of multimodal sentiment analysis in Arabic contexts. Unlike prior work that reports performance improvements in isolation, we argue that these challenges jointly explain the observed limitations across models and datasets.

1. **Morphological Richness and Dialectal Fragmentation:** Arabic is characterized by complex morphological structures, including rich inflection, clitic attachment, and non-concatenative word formation. This complexity introduces sparsity in lexical representations and increases ambiguity in sentiment interpretation. Moreover, the coexistence of Modern Standard Arabic (MSA) and numerous dialects (e.g., Egyptian, Gulf, Levantine) leads to substantial lexical and semantic variation. Models trained predominantly on MSA often struggle to generalize to dialectal expressions, resulting in degraded performance in real-world settings. Although pretrained models such as MarBERT partially address this issue by incorporating

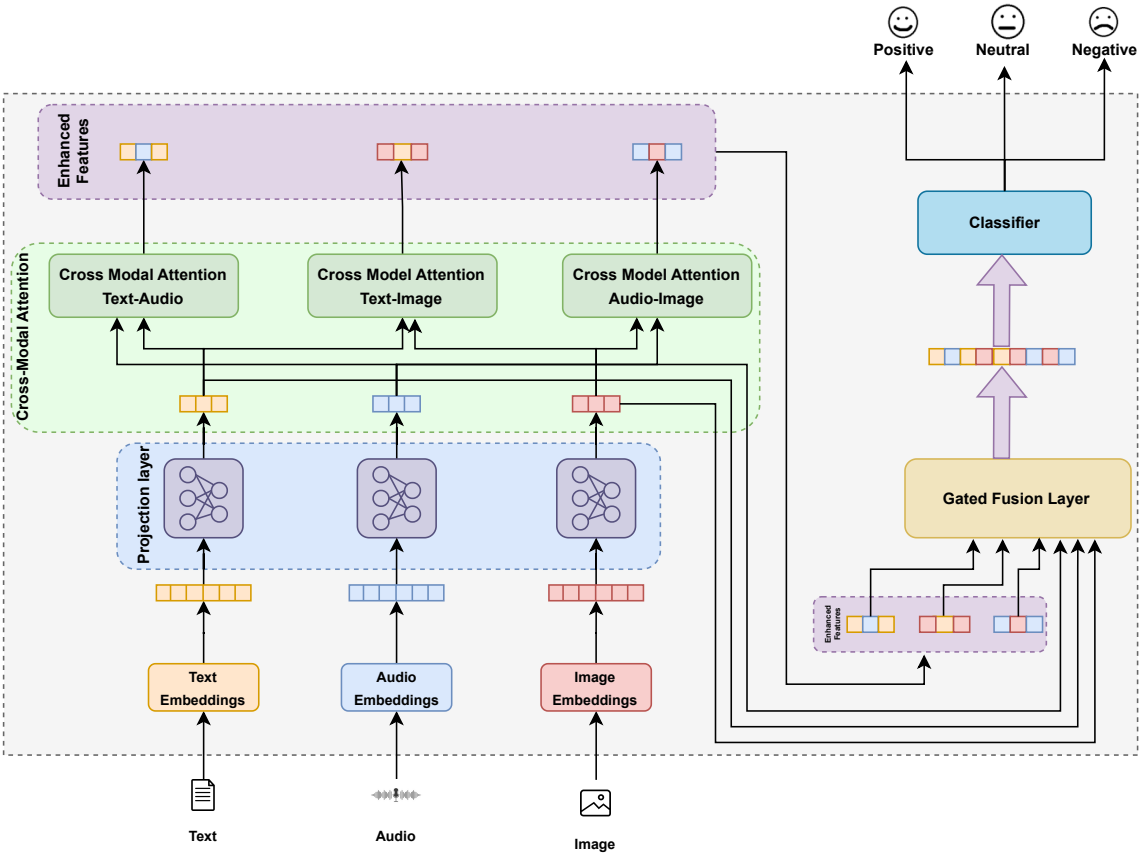

Figure 4. GCMA-Net architecture with cross-modal attention and gated fusion mechanism [39].

dialectal data, their effectiveness remains highly dependent on dataset composition [27, 46]. The impact of dialectal imbalance is evident in Ar-MuSA [27], where the dominance of Egyptian Arabic and MSA limits generalization to underrepresented dialects. Consequently, textual representations, while strong, are not uniformly robust across linguistic variations.

2. **Modality Reliability Imbalance:** A consistent finding across all studies is that textual modality significantly outperforms audio and visual modalities. For instance, in Ar-MuSA [27], text-based models achieve 71% F1-score, compared to 39% for audio and 45% for images. Similarly, in [46], text-only performance is comparable to or even exceeds multimodal configurations. This imbalance can be attributed to several factors. First, sentiment is often explicitly encoded in language through lexical cues, whereas audio and visual signals convey sentiment more implicitly through tone, prosody, or facial expressions. Second, pretrained models for Arabic text (e.g., MarBERT) are more mature and domain-adapted than their audio (HuBERT) and visual (ViT) counterparts, which are typically trained on non-Arabic or general-domain data. Furthermore, audio and visual modalities are more sensitive to noise and contextual variability, which reduces their reliability as standalone predictors. As a result, multimodal fusion often yields only marginal improvements over strong text-based baselines [27].
3. **Cross-Modal Misalignment:** Another critical challenge is the weak semantic alignment between modalities, particularly between textual and visual signals. In many real-world Arabic video datasets, visual content does not consistently correspond to the spoken or textual information, limiting its contribution to sentiment

prediction. This issue is explicitly observed in Ar-MuSA [27], where image-text fusion achieves only 46% F1-score, indicating limited complementary information. Similarly, early multimodal work [43] reports that audio-only models can outperform multimodal configurations, suggesting that poorly aligned modalities may introduce noise rather than useful signals. Naive fusion strategies, such as feature concatenation or majority voting, are insufficient to resolve such inconsistencies. More advanced approaches, such as UniTextFusion [45], attempt to mitigate this issue by converting all modalities into a unified textual representation. While this improves alignment, it may result in the loss of fine-grained acoustic and visual information. In contrast, interaction-aware architectures such as GCMA-Net [39] explicitly model cross-modal dependencies using attention mechanisms, achieving improved performance, although gains remain constrained by the underlying modality quality.

Overall, this taxonomy highlights that the limitations observed across existing studies are not solely due to architectural design choices, but rather stem from deeper linguistic, representational, and data alignment challenges inherent to Arabic multimodal settings.

6.2.2. Why Does Text Dominate? Across all surveyed studies, the textual modality consistently outperforms audio and visual modalities, often serving as the primary driver of overall system performance. For instance, in Ar-MuSA [27], text-based models achieve 71% F1-score, substantially exceeding audio (39%) and image (45%). A similar trend is observed in [46], where text-only performance (75.86% accuracy) rivals or even surpasses certain multimodal configurations. These findings suggest that multimodal gains are often bounded by the strength of the textual representation rather than the complementary contribution of additional modalities.

This dominance can be attributed to three interrelated factors:

- **Maturity and Domain Adaptation of Pretrained Language Models:** Textual representations benefit from highly optimized pretrained language models such as MarBERT, which are specifically trained on large-scale Arabic corpora, including dialectal data. These models capture nuanced semantic and contextual relationships, enabling robust sentiment detection even in noisy or informal text. In contrast, audio (e.g., HuBERT) and visual (e.g., ViT, MobileNetV2) encoders are typically pretrained on non-Arabic or general-domain datasets, limiting their ability to capture culturally and linguistically specific sentiment cues [27, 46]. As a result, the representational quality of non-text modalities is inherently weaker in Arabic contexts.
- **Explicit vs. Implicit Sentiment Encoding:** Sentiment is often directly and explicitly encoded in textual data through lexical and syntactic structures (e.g., polarity-bearing words, intensifiers, negation). For example, a sentence such as *"The movie was absolutely fantastic"* clearly conveys positive sentiment through explicit lexical markers. In contrast, audio and visual modalities encode sentiment more implicitly through prosody, tone, facial expressions, or gestures, which are inherently more ambiguous and context-dependent. This ambiguity makes it more difficult for models to reliably extract sentiment signals from these modalities, particularly in unconstrained environments such as user-generated video content.
- **Sensitivity to Noise and Contextual Variability:** Audio and visual modalities are significantly more sensitive to environmental noise and recording conditions. For instance, background noise, overlapping speech, or low-quality recordings can degrade the effectiveness of audio feature extraction. Similarly, visual signals in real-world datasets (e.g., YouTube videos in Ar-MuSA [27]) often lack clear facial expressions or are weakly correlated with the spoken content. This results in unstable or misleading features that reduce the overall reliability of these modalities. Empirical evidence supports this observation: in [27], audio-only performance is particularly low (39% F1), indicating that acoustic features alone are insufficient for robust sentiment prediction. Likewise, [43] reports cases where unimodal configurations outperform multimodal fusion due to noisy or weak visual features. These findings suggest that, rather than providing complementary information, additional modalities may introduce noise when not properly aligned or modeled.

Consequently, multimodal fusion strategies often yield only marginal improvements over strong text baselines. For example, UniTextFusion [45] achieves up to 71% F1-score—comparable to text-only performance—despite incorporating audio and visual information. This indicates that current multimodal approaches are limited not only

by architectural design but also by the intrinsic imbalance in modality informativeness. Overall, these observations highlight that the dominance of text is not incidental but structural, stemming from both the nature of sentiment expression and the relative maturity of textual modeling in Arabic. Addressing this imbalance requires not only improved fusion mechanisms but also stronger modality-specific representations for audio and visual data.

7. Fusion Mechanisms Under Arabic Constraints

A key limitation in prior work is the implicit assumption that improvements in multimodal sentiment analysis are primarily driven by architectural design. However, a closer examination of existing studies reveals that fusion effectiveness is strongly *data- and modality-dependent*, particularly in Arabic contexts where modality imbalance and weak cross-modal alignment are prevalent.

7.1. Early Fusion (Concatenation-Based)

Early fusion methods, such as the transformer-based framework proposed by [46], concatenate embeddings from MarBERT (text), HuBERT (audio), and ViT (image) into a unified representation (Figure 3). While this approach achieves competitive performance (77.6% F1-score), it relies on a strong assumption that modality-specific feature spaces are semantically aligned and equally informative. In practice, this assumption does not hold in Arabic multimodal settings. First, audio representations often fail to capture dialect-specific prosodic variations, which are critical for sentiment expression. This is reflected in the low standalone performance of audio models in Ar-MuSA (39% F1) [27]. Second, visual features extracted from unconstrained video data are frequently weakly correlated with sentiment. For example, a speaker expressing dissatisfaction may exhibit a neutral facial expression, resulting in conflicting signals between modalities. In such cases, simple concatenation introduces noise into the joint representation rather than complementary information. Consequently, early fusion models may overfit to dominant modalities (typically text) while underutilizing weaker ones.

7.2. Late Fusion (Voting and Ensembles)

Late fusion approaches, as implemented in Ar-MuSA [27], combine independent modality-specific predictions using voting or weighted ensembles. This strategy improves robustness by avoiding direct feature interaction; however, it assumes that each modality contributes reliable and complementary information. Empirical results challenge this assumption: the tri-modal voting ensemble achieves only 60% F1-score, which is lower than text-only performance. This outcome indicates that independent learning cannot compensate for weak or noisy modalities. For instance, if the textual modality confidently predicts a negative sentiment while the audio modality is uncertain due to background noise, a voting mechanism may dilute the correct prediction rather than reinforce it. Thus, late fusion fails to resolve modality conflicts and often underperforms when modality quality is imbalanced.

7.3. Feature/Score Hybrid Fusion

The hybrid fusion framework proposed by [44] (Figure 1) combines feature-level and score-level fusion strategies, achieving 94.08% accuracy on the SADAM dataset. While this result appears to significantly outperform other approaches, it must be interpreted in the context of dataset characteristics. The SADAM dataset is relatively small (524 utterances) and curated, which reduces noise and increases modality alignment. Notably, the visual modality achieves the highest unimodal performance (88% F1), in contrast to findings in Ar-MuSA [27], where visual signals are substantially weaker. This discrepancy highlights a critical insight: fusion performance is highly sensitive to dataset composition and modality quality. In controlled environments, modalities may appear strongly complementary; however, in real-world datasets, their contribution becomes inconsistent. Therefore, the effectiveness of hybrid fusion strategies is not universally transferable.

7.4. Textualized Fusion

UniTextFusion [45] (Figure 2) addresses cross-modal misalignment by converting audio and visual inputs into natural language descriptions, which are then fused with textual transcripts. This approach simplifies multimodal learning by leveraging the strength of large language models and achieves up to 71% F1-score. However, this design introduces a structural limitation: the transformation of non-textual modalities into text inevitably leads to information loss. Fine-grained acoustic features such as pitch variation, speech rate, or emphasis, as well as subtle visual cues like micro-expressions, are difficult to represent accurately in textual form. For example, sarcasm often relies on tonal shifts that cannot be fully captured through descriptive text. As a result, the model's performance remains bounded by the expressive capacity of the textual representation, limiting the potential gains from multimodal integration.

7.5. Interaction-Aware Fusion

Recent advances, such as GCMA-Net [39] (Figure 4), attempt to address these limitations through explicit cross-modal interaction modeling. By projecting modality-specific embeddings into a shared latent space and applying cross-modal attention, the model captures interdependencies between modalities. The addition of gated fusion further allows the model to dynamically weight modality contributions based on their reliability. This approach achieves 0.8015 F1-score on Ar-MuSA, outperforming early and late fusion baselines. Ablation studies reported in [39] demonstrate that removing cross-modal attention or gating mechanisms leads to measurable performance degradation, confirming their effectiveness. Nevertheless, the magnitude of improvement over early fusion methods remains relatively modest (approximately 2–3%), suggesting diminishing returns under current data conditions. Importantly, performance remains constrained by the inherent weakness of certain modalities, particularly audio. Even with sophisticated interaction modeling, the model cannot fully compensate for low-quality or weakly informative inputs. This indicates that architectural advancements alone are insufficient; improvements in modality representation and data quality are equally critical.

Overall, this analysis demonstrates that fusion effectiveness in Arabic multimodal sentiment analysis is not solely determined by model architecture. Instead, it is fundamentally shaped by modality reliability, dataset characteristics, and the degree of cross-modal alignment. Consequently, future research should focus not only on designing more complex fusion mechanisms but also on addressing the underlying data and representation challenges that limit multimodal learning.

8. Research Gaps

Despite the growing body of work on Arabic multimodal sentiment Analysis , the literature reviewed in Sections III and IV reveals several persistent structural and methodological gaps that hinder the development of robust and generalizable multimodal sentiment systems.

8.1. Limited Scale and Standardization of Multimodal Resources

As summarized in Table I, Arabic multimodal datasets remain relatively small, domain-specific, and unevenly accessible. While recent benchmarks such as Ar-MuSA and MSAD–Darija represent meaningful progress, their scale remains modest compared to widely used English multimodal corpora. Many earlier datasets are either unavailable or restricted in scope, limiting reproducibility and cross-study comparability. Furthermore, annotation practices vary considerably across datasets, particularly in how multimodal coherence is defined and evaluated. This fragmentation constrains the ability to train large joint representation models and prevents systematic benchmarking across dialects and domains.

8.2. Dialectal Diversity and Cross-Domain Generalization

Arabic is characterized by extensive dialectal variation, yet most multimodal sentiment datasets focus on a single dialect (e.g., Egyptian or Moroccan Arabic). Cross-dialect generalization remains largely unexplored. Models

trained on one dialect may not transfer effectively to others due to phonetic, lexical, and syntactic divergence. Moreover, multimodal sentiment cues may differ across cultural and regional contexts, further complicating generalization. Current research largely evaluates models within single-dataset settings, leaving cross-domain robustness insufficiently studied.

8.3. Persistent Modality Imbalance

Empirical evidence presented in Section IV indicates a consistent dominance of textual modality over audio and visual streams. Even in tri-modal configurations, text contributes the majority of predictive power, while acoustic and visual modalities provide comparatively weaker yet complementary signals. This imbalance suggests that existing fusion strategies do not fully exploit cross-modal dependencies. The problem is further exacerbated by noisy audio conditions, visually neutral frames, and transcription inaccuracies, particularly in user-generated content.

8.4. Fusion Strategy Limitations

Fusion design remains a central challenge. Early fusion methods based on embedding concatenation are computationally efficient but fail to explicitly model fine-grained inter-modal interactions. Late fusion and voting-based approaches improve modularity but do not capture deep contextual alignment. Although interaction-aware architectures such as gated cross-modal attention networks demonstrate improved performance, they introduce higher computational complexity and depend heavily on pretrained unimodal encoders. There remains a gap between scalable efficiency and expressive cross-modal reasoning.

8.5. Efficiency and Real-World Applicability

Computational efficiency remains an under-addressed challenge in Arabic multimodal sentiment analysis. Recent studies frequently rely on large-scale pretrained models with hundreds of millions or billions of parameters for relatively small sentiment datasets, which raises concerns regarding scientific proportionality and practical sustainability. Future research should prioritize Parameter-Efficient Fine-Tuning (PEFT) methods—such as adapters, LoRA, and prompt tuning—as well as knowledge distillation to develop lightweight yet competitive models. Moreover, real-world deployment constraints must be considered. In many Arabic-speaking regions, multimodal systems are expected to operate on mobile or edge devices with limited computational resources. Therefore, future work should incorporate efficiency-aware evaluation criteria, including model size and inference cost, to ensure both scalability and practical applicability.

9. Future Research Directions

Future studies should be required to (i) report performance under controlled dialectal splits to assess cross-dialect generalization, In particular, we advocate for the development of a large-scale, publicly available, multi-dialect Arabic multimodal benchmark with standardized annotation protocols and stratified sampling across dialects, explicitly designed to evaluate cross-dialect generalization, (ii) include robustness probes that isolate cultural and pragmatic expressions, and (iii) provide efficiency-aware metrics such as inference latency, model size, and energy footprint in addition to Accuracy and F1-score. Moreover, modality-level ablation must be systematically reported to quantify the marginal contribution of audio and visual features when text is present. Second, dataset development should move beyond syntactic annotation toward pragmatic and sociolinguistic representation. We recommend interdisciplinary collaboration with linguists and sociologists to construct corpora that capture discourse context, figurative language (e.g., irony and Tawriyah), and culturally grounded gestures. Future models should explicitly incorporate sociolinguistic variables—such as age, region, and gender—not merely as confounding factors but as structured features enabling socially aware representation learning. Such datasets should ensure balanced representation across dialects and modalities, enabling systematic analysis of dialectal variation and its interaction with multimodal signals. Third, modeling efforts should prioritize scalable and culturally grounded multimodal

architectures. This includes parameter-efficient adaptation, knowledge-aware semantic alignment, and robustness under noisy or incomplete modalities. In particular, future work should explore parameter-efficient fine-tuning (PEFT) approaches, such as adapter layers, to adapt multilingual multimodal foundation models to Arabic dialects while minimizing computational cost and explicitly modeling dialect-specific features across text, audio, and vision. Ultimately, the development of Arabic-centered multimodal foundation models jointly pretrained across text, speech, and vision—while adhering to standardized evaluation and deployment constraints—represents a critical milestone toward building trustworthy and socially contextualized multimodal AI systems.

10. Conclusion

This paper provided a comprehensive review of Multimodal Arabic Sentiment Analysis, study to analyze datasets, modeling strategies, and fusion paradigms. The survey highlighted the transition from handcrafted feature-based systems to transformer-driven and interaction-aware architectures integrating text, audio, and vision. Despite recent progress, particularly with the emergence of benchmarks such as Ar-MuSA and cross-modal attention models, Arabic multimodal research remains constrained by limited dataset scale, dialectal fragmentation, modality imbalance, and computational trade-offs in fusion design.

Overall, advancing Arabic multimodal sentiment analysis requires larger standardized resources, dialect-aware modeling, efficient interaction-centric architectures, and robust evaluation practices. Continued efforts toward unified multimodal pretraining and scalable Arabic-centered foundation models will be essential for developing accurate, reliable, and culturally grounded intelligent systems.

REFERENCES

1. A. B. Cheikhi *et al.*, "Multimodal arabic sarcasm detection using cnn and bilstm," in *2025 International Conference on Intelligent Systems: Theories and Applications (SITA)*. IEEE, 2025, pp. 1–5.
2. A. B. Cheikhi and E. H. Nfaoui, "Mathematics problem classification based on pretrained language models," in *2025 11th International Conference on Optimization and Applications (ICOA)*. IEEE, 2025, pp. 1–6.
3. N. Braig, A. Benz, S. Voth, J. Breitenbach, and R. Buettner, "Machine learning techniques for sentiment analysis of covid-19-related twitter data," *IEEE access*, vol. 11, pp. 14 778–14 803, 2023.
4. O. Elbiach, H. Grissette *et al.*, "Benchmarking large language models for adverse drug reaction extraction in social media and clinical texts," *Results in Engineering*, p. 107362, 2025.
5. R. Chandra, B. Zhu, Q. Fang, and E. Shinjikashvili, "Large language models for newspaper sentiment analysis during covid-19: The guardian," *Applied soft computing*, vol. 171, p. 112743, 2025.
6. H. Du, Q. Yu, and J. Chen, "Research on sentiment analysis of online public opinion based on multimodal big language modeling," *Journal of Computational Methods in Sciences and Engineering*, p. 14727978251361417, 2025.
7. C. Liu, M. Aljunied, G. Chen, H. P. Chan, W. Xu, Y. Rong, and W. Zhang, "Seallms-audio: Large audio-language models for southeast asia," *arXiv preprint arXiv:2511.01670*, 2025.
8. S. Lai, X. Hu, H. Xu, Z. Ren, and Z. Liu, "Multimodal sentiment analysis: A survey," *Displays*, vol. 80, p. 102563, 2023.
9. R. A. Stine, "Sentiment analysis," *Annual review of statistics and its application*, vol. 6, no. 1, pp. 287–308, 2019.
10. A. R. Sajun, I. Zualkernan, and D. Sankalpa, "A historical survey of advances in transformer architectures," *Applied Sciences*, vol. 14, no. 10, p. 4316, 2024.
11. J. Lehman, J. Gordon, S. Jain, K. Ndousse, C. Yeh, and K. O. Stanley, "Evolution through large models," in *Handbook of evolutionary machine learning*. Springer, 2023, pp. 331–366.
12. S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," *National Science Review*, vol. 11, no. 12, p. nwae403, 2024.
13. Z. Li, X. Wu, H. Du, F. Liu, H. Nghiem, and G. Shi, "A survey of state of the art large vision language models: Benchmark evaluations and challenges," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1587–1606.
14. B. Maji, M. Swain, S. Nasreen, D. Majumdar, R. Guha, A. Routray, and A. Søgaard, "A study on the impact of foundation models on automatic depression detection from speech signals," in *Proc. Interspeech 2025*, 2025, pp. 5258–5262.
15. A. AbdelRaouf, C. A. Higgins, T. Pridmore, and M. Khalil, "Building a multi-modal arabic corpus (mmac)," *International Journal on Document Analysis and Recognition*, vol. 13, no. 4, pp. 285–302, 2010.
16. S. Antar, A. Sagheer, S. Aly, and M. F. Tolba, "Avas: Speech database for multimodal recognition applications," in *13th International Conference on Hybrid Intelligent Systems (HIS)*. IEEE, 2013, pp. 123–128.
17. F. A. Shaqra, R. Duwairi, and M. Al-Ayyoub, "The audio-visual arabic dataset for natural emotions," in *Proceedings of the 7th International Conference on Future Internet of Things and Cloud (FiCloud)*. IEEE, 2019, pp. 324–329.
18. N. Al Roken and G. Barlas, "Multimodal arabic emotion recognition using deep learning," *Speech Communication*, vol. 155, p. 103005, 2023.

19. A. S. Alqarafi, A. Adeel, M. Gogate, K. Dashitpour, A. Hussain, and T. Durrani, "Toward's arabic multi-modal sentiment analysis," in *International Conference in Communications, Signal Processing, and Systems*. Springer, 2017, pp. 2378–2386.
20. R. M. Albalawi, A. T. Jamal, A. O. Khadidos, and A. M. Alhothali, "Multimodal arabic rumors detection," *IEEE Access*, vol. 11, pp. 9716–9730, 2023.
21. A. Hussein, M. Hussien, A. Hassona, W. Minker, M. A.-M. Salem, and N. Sharaf, "Egy-mer: Establishing the first egyptian arabic multimodal emotion recognition dataset for affective computing," 2025.
22. S. Ouali and S. El Garouani, "Mder-ma: A multimodal dataset for emotion recognition in low-resource moroccan arabic language," *Data in Brief*, p. 112005, 2025.
23. M. L. Bellagha and M. Zrigui, "Using the mgb-2 challenge data for creating a new multimodal dataset for speaker role recognition in arabic tv broadcasts," *Procedia Computer Science*, vol. 192, pp. 59–68, 2021.
24. H. Luqman, "Arabsign: A multi-modality dataset and benchmark for continuous arabic sign language recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2023, pp. 1–8.
25. S. Abbas, D. Alahmadi, and H. Al-Barhamtoshy, "Establishing a multimodal dataset for arabic sign language (arsl) production," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 8, p. 102165, 2024.
26. N. Alrashoudi, H. Al-Khalifa, and Y. Alotaibi, "Improving mispronunciation detection and diagnosis for non-native learners of the arabic language," *Discover Computing*, vol. 28, no. 1, p. 1, 2025.
27. S. Khaled, M. E. Ragab, A. K. Helmy, W. Medhat, and E. H. Mohamed, "Ar-musa: A multimodal benchmark dataset and evaluation framework for arabic sentiment analysis," *International Journal of Intelligent Engineering and Systems*, vol. 18, no. 4, 2025.
28. B. C. Ayoub and E. H. Nfaoui, "Msa-moroccan dialect: A multimodal sentiment analysis dataset for moroccan arabic (darija)," 2026. [Online]. Available: https://data.mendeley.com/datasets/kfjztyzztb
29. S. Ghaboura, A. Heakl, O. Thawakar, A. H. S. A. Alharthi, I. Riahi, A. Radman, J. Laaksonen, F. S. Khan, S. Khan, and R. M. Anwer, "Camel-bench: A comprehensive arabic lmm benchmark," in *Findings of the Association for Computational Linguistics: NAACL*, 2025, pp. 1970–1980.
30. A. Haouhat, S. Bellaouar, A. Nehar, and H. Cherroun, "Towards arabic multimodal dataset for sentiment analysis," in *Fourth International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*. IEEE, 2023, pp. 126–133.
31. C.-D. Nguyen, T. Nguyen, D. Vu, and L. A. Tuan, "Improving multimodal sentiment analysis: Supervised angular margin-based contrastive learning for enhanced fusion representation," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 14 714–14 724.
32. Z. Liu, B. Zhou, D. Chu, Y. Sun, and L. Meng, "Modality translation-based multimodal sentiment analysis under uncertain missing modalities," *Information Fusion*, vol. 101, p. 101973, 2024.
33. G. Paraskevopoulos, E. Georgiou, and A. Potamianos, "Mmlatch: Bottom-up top-down fusion for multimodal sentiment analysis," in *ICASSP*, 2022, pp. 4573–4577.
34. A. Zadeh *et al.*, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint arXiv:1707.07250*, 2017.
35. S. Mai *et al.*, "Hybrid contrastive learning of tri-modal representation," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 2276–2289, 2022.
36. H. Sun *et al.*, "Tensorformer: A tensor-based multimodal transformer," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 2776–2786, 2022.
37. F. Qian *et al.*, "Sentiment knowledge enhanced self-supervised learning," in *Findings of ACL*, 2023, pp. 12 966–12 978.
38. L. Christ *et al.*, "The muse 2023 multimodal sentiment analysis challenge," in *MuSe Workshop*, 2023, pp. 1–10.
39. A. B. Cheikhi and E. H. Nfaoui, "Gcma-net: A gated cross-modal attention network for arabic multimodal sentiment analysis," *IEEE Access*, 2026.
40. J. Zeng *et al.*, "Tag-assisted multimodal sentiment analysis," in *SIGIR*, 2022, pp. 1545–1554.
41. Y. Hwang and J.-H. Kim, "Self-supervised unimodal label generation strategy," in *Findings of ACL: EACL*, 2023, pp. 35–46.
42. Y. Zhao *et al.*, "Multimodal sentiment system and method based on crnn-svm," *Neural Computing and Applications*, vol. 35, no. 35, pp. 24 713–24 725, 2023.
43. H. Najadat and F. Abushaqra, "Multimodal sentiment analysis of arabic videos," *Journal of Image and Graphics*, vol. 6, no. 1, pp. 39–43, 2018.
44. S. Al-Azani and E.-S. M. El-Alfy, "Enhanced video analytics for sentiment analysis based on fusing textual, auditory and visual information," *IEEE Access*, vol. 8, pp. 136 843–136 857, 2020.
45. S. Khaled, W. Medhat, and E. H. Mohamed, "Unitextfusion: A low-resource framework for arabic multimodal sentiment analysis using early fusion and lora-tuned language models," *Ain Shams Engineering Journal*, vol. 16, no. 11, p. 103682, 2025.
46. A. B. CHEIKHI and E. H. NFAOUI, "A transformer-based approach for multimodal arabic sentiment analysis," *International Journal of Advanced Computer Science and Applications*, vol. 17, no. 2, 2026. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2026.0170290