machine learning & knowledge extraction

MDPI

*Article*

# A Directional Semantic Enhancement Approach with Gated Fusion for Multimodal Arabic Sentiment Analysis

**Ayoub Ben Cheikhi *, El Habib Nfaoui and Oumayma Elbiach**

L3IA Laboratory, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez 30000, Morocco; elhabib.nfaoui@usmba.ac.ma (E.H.N.); oumayma.elbiach@usmba.ac.ma (O.E.)
* Correspondence: ayoub.bencheikhi@usmba.ac.ma

**Abstract**

Multimodal Arabic sentiment analysis has gained increasing attention due to the growing volume of user-generated multimedia content. However, effectively integrating textual, acoustic, and visual modalities remains challenging because of modality imbalance and weak cross-modal alignment. This study proposes a Directional Semantic Enhancement approach with Gated Fusion to address these limitations. The objective is to explicitly model similarity-guided semantic transfer between modalities while adaptively regulating information flow during fusion. The proposed architecture consists of four main stages: modality encoding, directional semantic enhancement, gated fusion, and classification. Directional semantic interactions enable structured cross-modal knowledge exchange, while adaptive gating mechanisms balance original and enhanced representations to mitigate modality-specific noise. Extensive experiments are conducted on the Ar-MuSA benchmark dataset, which contains 8700 multimodal samples. The proposed approach achieves 89.89% accuracy and an F1-score of 0.8989 with a latent dimension of 1024, outperforming early fusion, late fusion, and recent state-of-the-art methods. The study highlights the importance of controlled cross-modal alignment and provides a scalable approach for robust multimodal sentiment understanding in Arabic multimedia environments.

**Keywords:** Arabic multimodal sentiment analysis; directional semantic enhancement; gated fusion; feature projection; deep learning

## 1. Introduction

Artificial intelligence has become an essential technological driver across a wide range of domains, including Archaeology [1], natural language processing applications [2–4]. In particular, Sentiment analysis has become a central task in natural language processing and affective computing, driven by the rapid growth of user-generated multimedia content across social media platforms, online video streams, and digital communication channels. While early research primarily focused on textual sentiment classification, real-world human communication rarely relies on text alone [5–7]. Instead, sentiment emerges from the interaction between linguistic meaning, vocal tone, facial expressions, and contextual visual cues. This observation has led to the development of Multimodal Sentiment Analysis (MuSA), which aims to integrate heterogeneous modalities such as text, audio, and visual information to achieve more accurate and robust affect recognition [8–11]. In the context of the Arabic language, multimodal sentiment analysis presents additional challenges [12–15]. Arabic is morphologically rich and dialectally diverse, with significant

Check for updates

Academic Editor: Jaroslaw Krzywanski

Received: 30 March 2026
Revised: 1 May 2026
Accepted: 18 May 2026
Published: 21 May 2026

**Copyright:** © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

variations across regions and informal communication styles. Furthermore, modality imbalance remains a critical issue in Arabic multimodal datasets, where textual features often dominate predictive performance while audio and visual signals contribute inconsistently. Existing fusion strategies typically rely on early feature concatenation or late decision aggregation [16,17]. However, such approaches often fail to explicitly model structured cross-modal dependencies and may suffer from noisy or unreliable modality interactions. As a result, effective cross-modal alignment and controlled information exchange remain open research problems in multimodal Arabic sentiment analysis.

To address these challenges, this paper proposes a Directional Semantic Enhancement approach with Gated Fusion specifically tailored for Multimodal Arabic Sentiment Analysis. The architecture is designed to bridge the research gap in effectively integrating heterogeneous modalities within the complex linguistic and dialectal landscape of the Arabic language. The proposed approach is designed to explicitly model similarity-guided semantic transfer between modalities while adaptively regulating the flow of information. Instead of treating modalities as independent contributors or simply concatenating their embeddings, the model introduces directional semantic interactions that allow each modality to receive complementary information from others in a controlled manner. A gating mechanism is then employed to balance original and enhanced representations, mitigating modality-specific noise and reducing cross-modal conflicts.

The main contributions of this work can be summarized as follows:

- We introduce a directional semantic enhancement mechanism that explicitly models similarity-aware cross-modal interactions between text, audio, and visual representations.
- We propose an adaptive gated fusion strategy that dynamically balances modality-specific and cross-modally enriched features.
- We conduct extensive experiments on the Ar-MuSA benchmark dataset, demonstrating significant improvements over baseline and state-of-the-art multimodal fusion approaches.

The remainder of this paper is organized as follows. In the next section, we review related work in Arabic multimodal sentiment analysis, multimodal fusion strategies, and direction-aware semantic modeling. Then, we present the proposed multimodal architecture, including modality encoding, directional semantic enhancement, and gated fusion mechanisms. After that, we describe the experimental setup, including dataset characteristics, preprocessing pipeline, training configuration, and evaluation metrics. Subsequently, we report and analyze the experimental results through overall performance comparison, class-wise evaluation, and ablation studies. Before concluding, we discuss the implications of our findings. Finally, we conclude the paper with directions for future research.

## 2. Related Work

This section provides a critical overview of existing research in multimodal sentiment analysis, with particular emphasis on Arabic-language settings. We first review prior approaches to Arabic multimodal sentiment analysis and multimodal fusion strategies. We then examine recent advances in direction-aware semantic modeling and interaction mechanisms. Finally, we identify the methodological limitations that motivate the proposed directional semantic enhancement with a gated fusion approach.

### *2.1. Multimodal Fusion Strategies*

Fusion techniques are generally categorized into early, late, hybrid, and attention-based strategies.

Early fusion integrates modality-specific embeddings before classification. Contrastive learning mechanisms have been proposed to enhance fusion representation quality while

preserving unimodal importance [18]. Translation-based approaches have also addressed uncertain missing modalities by projecting audio and visual features into textual representation spaces [19]. Late fusion combines independent modality predictions at the decision level. Hierarchical bottom-up and top-down fusion architectures incorporating cross-modal masking have achieved state-of-the-art performance on benchmark datasets [20]. Public multimodal sentiment challenges have further established evaluation standards and baseline comparisons [21]. Multimodal frameworks for Arabic sarcasm detection have also demonstrated the effectiveness of integrating CNN, BiLSTM, and deep visual encoders [22]. Hybrid fusion approaches combine feature-level integration with decision-level optimization. CRNN-SVM architectures have achieved robust multimodal sentiment recognition performance [23]. Tensor-based fusion networks explicitly model intra- and inter-modal interactions through higher-order representations [24]. Attention-based multimodal fusion has further advanced cross-modal reasoning capabilities. Temporal attention mechanisms have improved multimodal video understanding [25]. Mutual attention frameworks have enhanced multimodal fake news detection [26]. Contextual attention combining convolutional and large language models has strengthened multimodal sentiment detection in crisis-related data [27]. Structured self-attention architectures integrating vision transformers and language models have improved semantic alignment across modalities [28].

### 2.2. Arabic Multimodal Sentiment Analysis

Multimodal Sentiment Analysis (MuSA) has gained increasing attention in affective computing, particularly in morphologically rich and dialectally diverse languages such as Arabic [15,16,29]. The integration of textual, acoustic, and visual modalities has demonstrated clear advantages over unimodal systems in capturing implicit affective cues.

Early efforts in Arabic multimodal sentiment analysis relied on handcrafted feature engineering and traditional machine learning classifiers. A pioneering framework integrating linguistic, acoustic, and facial features extracted from Arabic YouTube videos demonstrated the feasibility of multimodal sentiment detection, achieving 76% accuracy [30]. However, shallow feature representations and basic feature concatenation limited the modeling of complex cross-modal interactions. Subsequent research introduced recurrent neural network architectures with attention-based fusion mechanisms to enhance inter-modal learning [31]. Although improvements were observed over unimodal baselines, validation was primarily conducted on non-Arabic datasets, leaving Arabic-specific linguistic and dialectal challenges insufficiently addressed. A more systematic evaluation of multimodal fusion strategies in Arabic was conducted through feature-level, score-level, and decision-level integration approaches [32]. While hybrid fusion achieved high reported accuracy, the limited dataset size and reliance on handcrafted descriptors constrained scalability and generalization across dialects. The challenge of Egyptian Arabic–English code-switching was addressed through the development of a dedicated sentiment corpus evaluated using both traditional neural models and large language models [33]. Only marginal performance gains were observed for fine-tuned transformer models compared to optimized neural baselines, suggesting that architectural complexity alone does not guarantee improved generalization in limited-resource settings. To standardize evaluation in Arabic MuSA, a larger aligned multimodal benchmark was introduced, comprising text, audio, and image samples extracted from Egyptian YouTube videos [34]. Experimental results revealed a dominance of textual modality, whereas audio and visual signals contributed weaker standalone performance. Late fusion provided moderate complementary gains but failed to fully exploit cross-modal synergy. A computationally efficient early fusion framework was later proposed by transforming acoustic and visual inputs into descriptive textual prompts compatible with large language models [17]. Although suitable for low-resource

environments, this modality textualization process may compress fine-grained affective information. Transformer-based multimodal architectures further explored early embedding concatenation combined with margin-based classifiers [16]. While performance was achieved, explicit modeling of cross-modal alignment remained limited. To address modality imbalance and weak interaction modeling, the authors [29] propose a gated cross-modal attention architecture incorporating projection layers, attention pathways, and reliability-aware fusion. This approach demonstrated statistically significant improvements over early and late fusion baselines, emphasizing the importance of interaction-aware multimodal learning in multimodal Arabic sentiment analysis. Despite notable advancements, several limitations persist in Arabic multimodal sentiment analysis. First, modality imbalance remains a fundamental issue, with textual features consistently dominating predictive performance, limiting effective utilization of audio and visual affective cues [34]. Second, early fusion methods based on simple embedding concatenation fail to explicitly capture fine-grained cross-modal dependencies [16]. Conversely, late fusion approaches treat modalities independently, restricting hierarchical semantic integration [20]. Third, available Arabic multimodal datasets remain relatively small compared to English benchmarks, limiting large-scale optimization and robustness evaluation [34].

Therefore, there remains a clear need for a structured, interaction-aware multimodal architecture that (i) explicitly models directional semantic transfer between modalities, (ii) employs similarity-guided gating to regulate cross-modal influence, and (iii) adaptively balances original and enhanced representations to mitigate modality-specific noise. These limitations directly motivate the proposed directional semantic enhancement with a gated fusion approach, which aims to provide controlled cross-modal interaction while preserving modality-specific discriminative information.

### *2.3. Directional Semantic Enhancement Applications*

In the context of multimodal sentiment analysis, Directional Semantic Enhancement refers to the guided and explicit transfer of affective information between modalities, aiming to enrich or disambiguate the semantic representation of a target modality. While the concept of 'directionality' has been extensively explored in spatial domains, such as remote sensing and visual grounding, its underlying principle of structured, guided information flow provides valuable theoretical inspiration for semantic interactions in sentiment analysis. The goal is to achieve affective localization—identifying and reinforcing sentiment-relevant cues—analogous to how spatial directionality aids in object localization.

Recent developments in multimodal learning have emphasized the importance of directional and spatial semantic modeling in strengthening cross-modal interaction [35,36]. Although most existing work has focused on natural image understanding and remote sensing tasks, the underlying principle of directional semantic enhancement offers valuable analogies and inspirations for multimodal sentiment analysis. For instance, in salient object detection for optical remote sensing imagery, directional spectral enhancement has been combined with hierarchical text–visual fusion to reinforce boundary representation and semantic alignment [37]. This illustrates how structured information flow, adapted to semantic cues, can refine multimodal representations beyond simple feature concatenation.

This highlights how structured spatial-semantic modeling can refine multimodal representations beyond simple feature concatenation. Similarly, direction-oriented visual–semantic embedding strategies have been proposed to reduce visual–semantic imbalance in cross-modal retrieval tasks [38]. By guiding latent embeddings toward region-aware representations and enforcing global semantic constraints, cross-modal alignment becomes more coherent and less sensitive to misleading visual similarities. Such mechanisms illustrate the importance of structured embedding alignment in multimodal interaction. In

supervised cross-modal hashing, multi-grained semantic fusion and bi-directional relational aggregation have been introduced to strengthen semantic distribution consistency across modalities [39]. This relational enhancement approach demonstrates that preserving structural dependencies between modalities leads to more discriminative and semantically aligned representations. Directional semantic modeling has also been applied to visual grounding tasks, where explicit spatial semantic alignment mechanisms enhance localization performance [40]. By injecting direction-aware prompts and scale-adaptive language guidance into decoder stages, models achieve improved cross-modal reasoning in spatially complex scenes. Although these approaches were developed for vision-centric applications, they collectively reveal a broader methodological shift toward structured, interaction-aware semantic modeling. For multimodal Arabic sentiment analysis, where modality imbalance and weak cross-modal alignment remain critical challenges, such direction-aware and relational semantic enhancement strategies provide conceptual foundations for designing more robust multimodal fusion architectures.

## 3. Proposed Approach

In this section, we present the proposed multimodal Approach for Arabic sentiment analysis classification, which aims to effectively integrate textual, acoustic, and visual information. The model is designed to capture both modality-specific characteristics and cross-modal interactions in a unified architecture. As illustrated in Figure 1, the proposed architecture consists of four main stages: modality encoding, directional semantic enhancement, gated fusion, and classification. These components are jointly optimized to produce robust and discriminative multimodal representations.

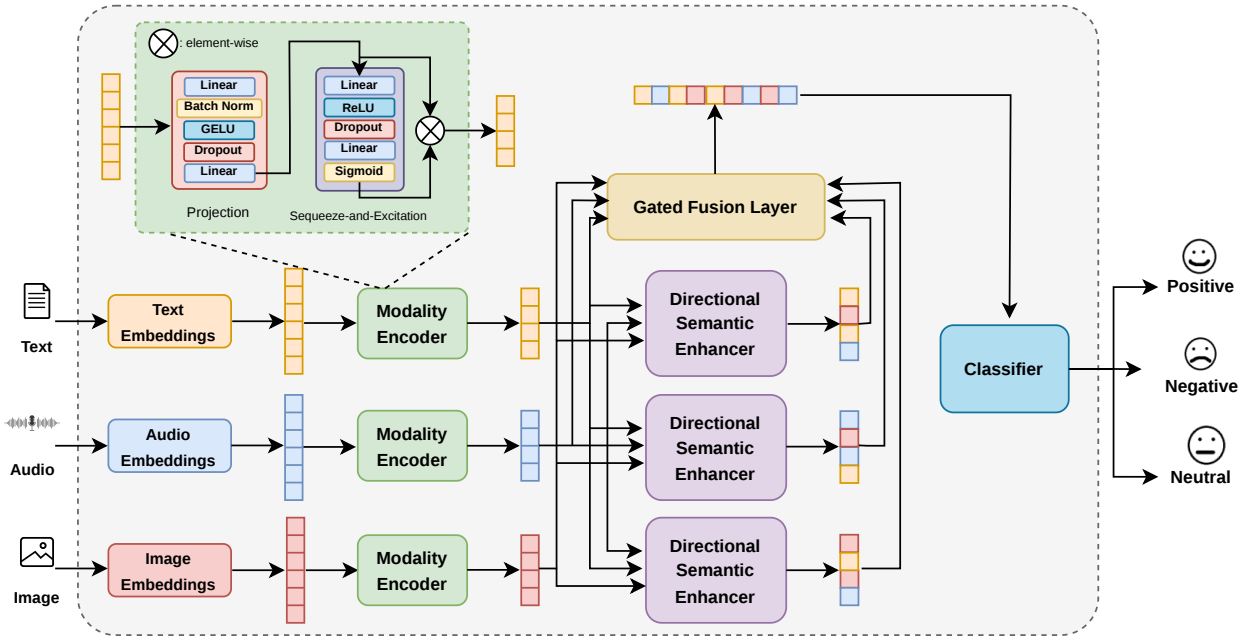



**Figure 1.** Overview of the proposed multimodal architecture. The model first encodes each modality independently, then captures cross-modal relationships through directional semantic enhancement, followed by gated fusion to adaptively combine representations. Finally, the fused features are used for sentiment classification.

### *3.1. Problem Formulation*

A multimodal instance (See the example illustrated in Figure 2) in our approach is represented as:

$$X = (T, A, V) \tag{1}$$

where $T \in \mathbb{R}^{d_t}$, $A \in \mathbb{R}^{d_a}$, and $V \in \mathbb{R}^{d_v}$ correspond to the feature representations of the text, audio, and visual modalities, respectively. These features are obtained from pretrained encoders and capture high-level semantic, acoustic, and visual information.

Given a dataset

$$\mathcal{D} = \{(T^{(i)}, A^{(i)}, V^{(i)}, y^{(i)})\}_{i=1}^{N} \tag{2}$$

where $y^{(i)} \in \{1, \ldots, C\}$ denotes the sentiment label, the objective of our approach is to learn a function:

$$f : (T, A, V) \rightarrow \hat{y} \tag{3}$$

that accurately predicts the sentiment class while remaining robust to modality-specific noise through effective multimodal fusion.




**Figure 2.** Example of an instance from the Ar-MuSA dataset.

3.1.1. Modality Encoding

In this stage, each modality is first transformed into a latent representation through a learnable projection. Specifically, for a given modality $m \in \{t, a, v\}$, the input features are mapped as:

$$\mathbf{h}^{(m)} = \text{GELU}(\text{BN}(W_m \mathbf{x}^{(m)} + b_m)) \tag{4}$$

where $W_m$ and $b_m$ denote learnable parameters, and $\text{BN}(\cdot)$ represents batch normalization. We adopt the GELU activation function owing to its smooth nonlinear characteristics, ability to preserve small informative signals compared to ReLU's hard thresholding and its demonstrated effectiveness in transformer-based representation learning [41]. Additionally, Batch Normalization and dropout are incorporated to stabilize the training process and enhance robustness to modality-specific noise and overfitting [42,43].

Next, a feature-wise attention mechanism, inspired by Squeeze-and-Excitation [44], is applied to emphasize informative components. The attention weights are computed as:

$$\mathbf{s}^{(m)} = \sigma(W_2 \, \text{ReLU}(W_1 \mathbf{h}^{(m)})) \tag{5}$$

where $\sigma$ denotes the sigmoid function.

This mechanism enables the model to assign higher importance to salient features while suppressing less relevant ones within each modality.

Finally, the refined modality representation is obtained via element-wise scaling:

$$\mathbf{z}^{(m)} = \mathbf{h}^{(m)} \odot \mathbf{s}^{(m)} \tag{6}$$

where $\odot$ denotes element-wise multiplication.

3.1.2. Directional Semantic Enhancement

After modality encoding, the model captures cross-modal relationships through directional semantic interactions, as illustrated in Figure 3. This mechanism allows each modality to incorporate relevant information from others while preserving its own characteristics.

The similarity between two modalities is computed using cosine similarity:

$$s_{m_1 m_2} = \frac{\mathbf{z}^{(m_1)} \cdot \mathbf{z}^{(m_2)}}{\|\mathbf{z}^{(m_1)}\| \|\mathbf{z}^{(m_2)}\|} \tag{7}$$

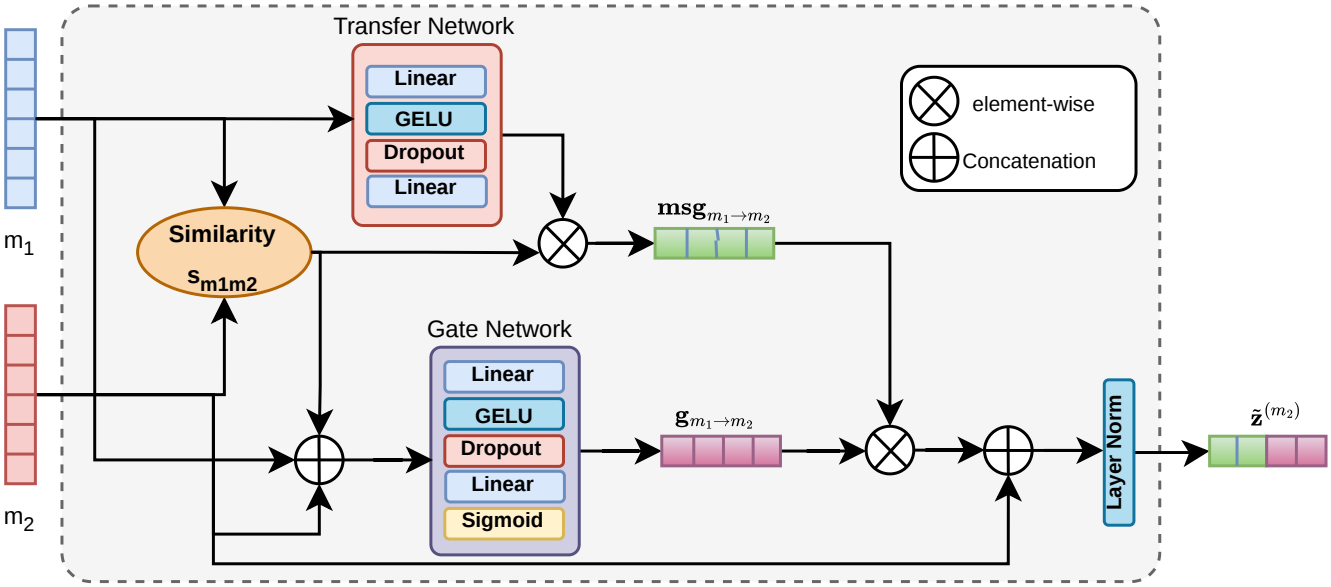



**Figure 3.** The architecture of the directional semantic interactions.

A semantic message from modality $m_1$ to modality $m_2$ is then defined as:

$$\mathbf{msg}_{m_1 \to m_2} = T(\mathbf{z}^{(m_1)}) \odot s_{m_1 m_2} \tag{8}$$

where $T(\cdot)$ is a learnable transformation. For example, when two modalities are aligned, the transferred information becomes more significant.

To control the contribution of this message, a gating mechanism is applied:

$$\mathbf{g}_{m_1 \to m_2} = \sigma(W_g[\mathbf{z}^{(m_1)}; \mathbf{z}^{(m_2)}; s_{m_1 m_2}]) \tag{9}$$

The enhanced representation of modality $m_2$ is computed as:

$$\tilde{\mathbf{z}}^{(m_2)} = \text{LayerNorm}(\mathbf{z}^{(m_2)} + \alpha\, \mathbf{g}_{m_1 \to m_2} \odot \mathbf{m}_{m_1 \to m_2}) \tag{10}$$

where $\alpha$ is a learnable scaling parameter.

Each modality receives information from the other two modalities, and the final enhanced representation is obtained by averaging:

$$\mathbf{e}^{(m)} = \frac{1}{2}(\tilde{\mathbf{z}}_1^{(m)} + \tilde{\mathbf{z}}_2^{(m)}) \tag{11}$$

This approach enables the model to capture meaningful cross-modal dependencies while remaining robust to noise.

3.1.3. Gated Fusion

After Directional Semantic Enhancement, we apply a gated fusion mechanism to each modality and its enhanced representation, as illustrated in Figure 4. This step aims to adaptively combine the original encoded features with the cross-modally enriched representations, allowing the model to balance complementary information while reducing the impact of noisy interactions.

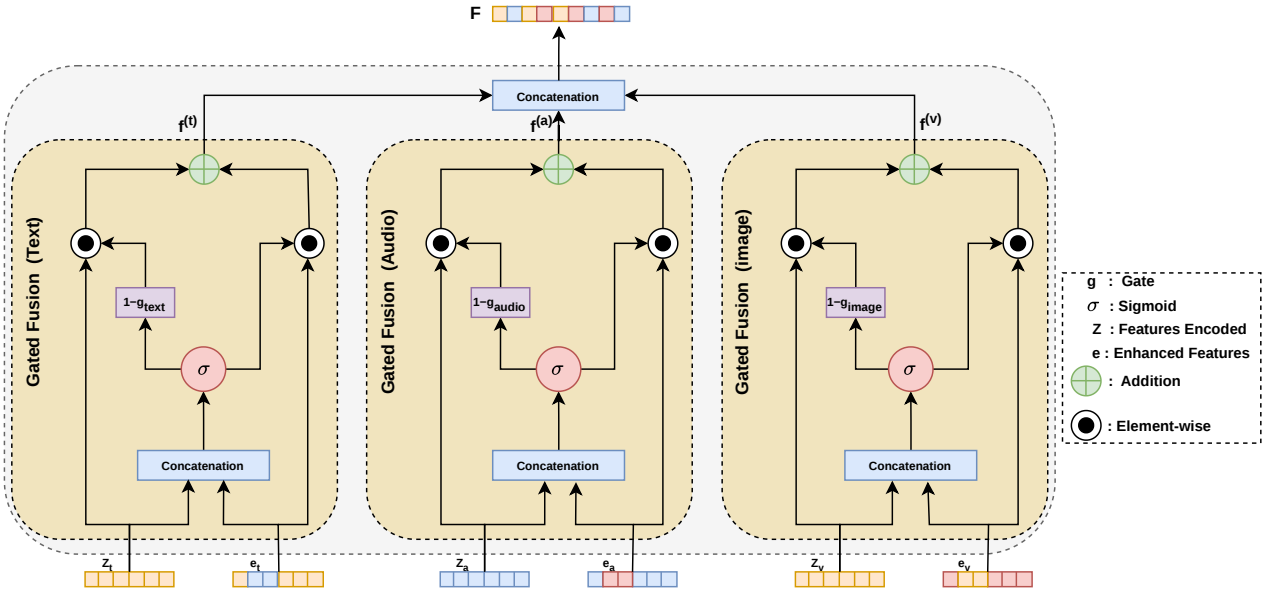



**Figure 4.** The architecture of the Gated Fusion.

First, a gating vector is computed as:

$$\mathbf{g}^{(m)} = \sigma(W_f[\mathbf{z}^{(m)}; \mathbf{e}^{(m)}]) \tag{12}$$

where $[\cdot;\cdot]$ denotes concatenation and $\sigma$ is the sigmoid activation.

The final representation of modality $m$ is then obtained by:

$$\mathbf{f}^{(m)} = \mathbf{g}^{(m)} \odot \mathbf{e}^{(m)} + (1 - \mathbf{g}^{(m)}) \odot \mathbf{z}^{(m)} \tag{13}$$

This formulation enables a soft and adaptive selection between the original representation $\mathbf{z}^{(m)}$ and the enhanced representation $\mathbf{e}^{(m)}$. For example, when cross-modal interactions provide useful complementary information, the model assigns higher weights to $\mathbf{e}^{(m)}$; otherwise, it preserves more of the original modality-specific features. This behavior improves robustness, particularly in cases where some modalities may be noisy or less informative.

The use of gating mechanisms is motivated by their effectiveness in controlling information flow and handling uncertainty in neural networks. In the multimodal context, such adaptive fusion is essential due to the heterogeneous and potentially unreliable nature of different modalities.

Finally, a normalization layer is applied to stabilize the fused representation and facilitate training.

3.1.4. Classification

After gated fusion, the final representation of each modality is obtained as $\mathbf{f}^{(t)}$, $\mathbf{f}^{(a)}$, and $\mathbf{f}^{(v)}$. These modality-specific representations are concatenated to form the global multimodal feature vector:

$$\mathbf{F} = [\mathbf{f}^{(t)}; \mathbf{f}^{(a)}; \mathbf{f}^{(v)}] \tag{14}$$

where $[\cdot;\cdot]$ denotes vector concatenation.

This fused representation is then passed through a feed-forward classifier to produce a compact hidden representation:

$$\mathbf{h} = \text{Dropout}\Big(\text{GELU}\big(\text{LayerNorm}(W_c\mathbf{F} + b_c)\big)\Big) \tag{15}$$

where $W_c$ and $b_c$ are learnable parameters. The output logits are computed as:

$$\mathbf{o} = W_o\mathbf{h} + b_o \tag{16}$$

where $W_o$ and $b_o$ denote the parameters of the output layer. The predicted sentiment distribution is finally obtained by:

$$\hat{\mathbf{y}} = \text{Softmax}(\mathbf{o}) \tag{17}$$

where $\hat{\mathbf{y}} \in \mathbb{R}^C$ represents the probability distribution over the $C$ sentiment classes.

## 4. Experimental Setup

This section describes the experimental setup used to evaluate the proposed multimodal Arabic sentiment analysis approach. It includes a detailed description of the dataset, preprocessing pipeline, training configuration, and evaluation metrics. First, we introduce the Ar-MuSA dataset and analyze its statistical properties across modalities. Then, we present the preprocessing strategy used to extract text, audio, and visual embeddings. After that, we describe the training configuration of the proposed model. Finally, we outline the evaluation measures adopted to assess the performance from both overall and class-wise perspectives.

### *4.1. Datasets*

This subsection describes the dataset used in our experiments and provides an overview of its statistical properties. We conduct our evaluation on the Ar-MuSA corpus, which consists of 8700 synchronized multimodal samples extracted from 107 YouTube videos. Each sample includes a textual transcript, the corresponding audio segment, and an aligned visual frame, all captured at the same temporal instance. The Ar-MuSA dataset is publicly available and can be accessed at: https://huggingface.co/datasets/Skhaled/Ar-MUSA (accessed on 1 May 2026).

The annotation process involved an initial independent assessment of each modality without cross-referencing the others. Subsequently, a unified overall sentiment label—Positive, Negative, or Neutral—was assigned to each multimodal instance by considering the combined cues from all three modalities. Table 1 presents the distribution of these final sentiment labels in terms of both absolute counts and percentage ratios.

The Ar-MuSA dataset is organized into multiple annotated subsets (Annotation 1a, Annotation 1b, and Annotation 2a). Each subset contains four main components: (i) a frames directory including the extracted visual frames for each record, (ii) an audios directory containing the corresponding audio segments, (iii) a transcripts directory providing the textual transcriptions of the audio content, and (iv) an annotations.csv file that stores

the sentiment labels and metadata for each multimodal instance. Each sample is therefore defined by aligned textual, acoustic, and visual data linked through the annotation file, which specifies the sentiment category and record identifier.

**Table 1.** Sentiment Distribution Across Modalities.

| Modality | Positive | Negative | Neutral |
|---|---|---|---|
| Text | 3092 (35.5%) | 1944 (22.3%) | 3664 (42.1%) |
| Audio | 3151 (36.2%) | 1949 (22.4%) | 3600 (41.3%) |
| Image | 2015 (23.1%) | 1980 (22.7%) | 4705 (54.1%) |

As illustrated in Table 1 and Figure 5, the distribution of sentiment labels varies across modalities. For example, the visual modality contains a relatively higher proportion of Neutral samples, which often correspond to non-expressive or ambiguous frames. This imbalance highlights the inherent variability in modality reliability.

Although no explicit re-sampling strategy is applied, the proposed model addresses this issue through gated fusion and dropout-based regularization. These mechanisms enable the network to adaptively down-weight less informative modalities, thereby improving robustness during multimodal integration.

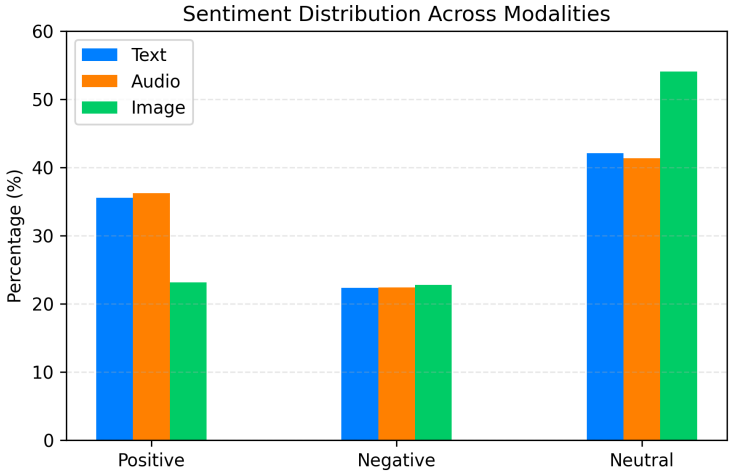



**Figure 5.** Distribution of sentiment labels across text, audio, and visual modalities in the ArMuSA dataset.

### *4.2. Preprocessing*

In this subsection, we describe the preprocessing pipeline used to obtain unified and informative representations for each modality. The goal is to transform heterogeneous multimodal inputs into feature embeddings prior to fusion.

**Text Embeddings.** For the textual modality, we employ MARBERT [45], a pretrained transformer model specifically designed for Arabic language understanding. It is trained on both Modern Standard Arabic and dialectal variations, enabling it to capture rich linguistic and contextual information. The transcripts are tokenized and processed using MARBERT to extract contextual embeddings with dimensions d_t = 768, that reflect semantic meaning and sentiment cues. The textual transcripts preserve sentiment-relevant symbolic elements such as punctuation marks (e.g., exclamation and question marks) and expressive markers. During preprocessing, these elements were retained without aggressive normalization to preserve affective cues before feature extraction with MARBERT.

**Audio Embeddings.** For the audio modality, we utilize HuBERT [46], a self-supervised speech representation model. Audio signals are first resampled and normalized to 16 kHz,

then passed through HuBERT to obtain acoustic embeddings with dimensions d_a = 1024. These representations encode important paralinguistic features such as prosody, intonation, and emotional tone, which are essential for sentiment analysis.

**Image Embeddings.** For the visual modality, we adopt the Vision Transformer (ViT) [47]. Extracted video frames are resized and normalized before being processed by the pretrained ViT model. The resulting embeddings with dimensions d_v = 768, capture high-level visual patterns, including facial expressions and other visual cues related to emotional states.

This preprocessing strategy ensures that all modalities are represented in a consistent and semantically rich embedding space. Although MARBERT, HuBERT, and ViT offer powerful pretrained representations, their computational complexity can be restrictive in certain scenarios. Therefore, in our implementation, embeddings are extracted offline prior to model training. This design choice allows the proposed approach to remain efficient during the fusion and inference stages while still benefiting from high-quality pretrained features.

### *4.3. Training Configuration*

In this subsection, we describe the training setup and implementation details of the proposed approach. The dataset is partitioned using Scikit-learn's train_test_split function with a stratified random split of 80% for training, 10% for validation, and 10% for testing, ensuring consistent class distribution across all subsets. The validation set was used for model selection based on weighted F1-score, with early stopping applied after 8 epochs without improvement. We also fixed the random seed to 42 for Python, NumPy, and PyTorch. The model is implemented using the PyTorch framework and trained end-to-end on multimodal embeddings extracted from MARBERT for text, HuBERT for audio, and ViT for image representations. The training configuration uses a batch size of 8, a latent dimensions (128, 256, 512, and 1024) to analyze the impact of representational capacity on performance, a dropout rate of 0.35, and a modality dropout of 0.15. The model is optimized using the AdamW optimizer with a learning rate of $7 \times 10^{-4}$ and a weight decay of $5 \times 10^{-4}$ for 30 epochs. Additionally, $\alpha$ is introduced as a learnable scaling parameter and initialized to 0.15. The model is trained using a weighted cross-entropy loss with label smoothing to address class imbalance and improve calibration. Furthermore, mixup is applied in the latent space, and a similarity regularization term is incorporated to enforce cross-modal consistency. A learning rate scheduler is employed to adaptively reduce the learning rate based on validation performance, while early stopping is used to prevent overfitting. This configuration ensures stable convergence and enables the proposed approach to effectively learn robust multimodal representations. All experiments were conducted in Python 3.12.13 using PyTorch 2.10.0 with CUDA 12.8 support. The model was trained on Google Colab utilizing an NVIDIA L4 GPU with 22 GB of GPU memory and approximately 53 GB of system RAM. CUDA acceleration was enabled to ensure efficient training and optimization.

### *4.4. Evaluation Measures*

In this subsection, we describe the evaluation metrics used to assess the performance of the proposed approach. Since no single metric can fully capture the behavior of a multimodal classification model, multiple complementary measures are considered to provide a comprehensive evaluation.

**Precision.** Precision evaluates the reliability of positive predictions by measuring the proportion of correctly predicted positive instances among all predicted positives. It is defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{18}$$

where $TP$ and $FP$ denote true positives and false positives, respectively. For example, high precision indicates that the model produces fewer false alarms.

**Recall.** Recall, also referred to as sensitivity, measures the model's ability to correctly identify all relevant instances. It is computed as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{19}$$

where $FN$ denotes false negatives. A high recall value indicates that most positive samples are successfully detected.

**F1-Score.** The F1-score provides a balanced measure by combining precision and recall through their harmonic mean:

$$\text{F1-score} = \frac{2TP}{2TP + FP + FN} \tag{20}$$

The F1-score is computed using standard macro-averaging across the three sentiment classes, assigning equal weight to each class regardless of class frequency. This metric is particularly useful when dealing with imbalanced datasets, as it balances both types of errors.

**Accuracy.** Accuracy measures the overall proportion of correctly classified instances:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{21}$$

where $TN$ denotes true negatives. Although widely used, accuracy alone may be misleading in the presence of class imbalance. This combination ensures both global and class-level behaviors are properly assessed.

## 5. Results

This section presents a comprehensive evaluation of the proposed directional semantic enhancement with a gated fusion approach on the Ar-MuSA dataset. The analysis is organized into four parts. First, we examine the impact of latent representation dimensionality on overall performance. Second, we compare the proposed approach with baseline and state-of-the-art multimodal methods. Third, we provide a detailed class-wise performance analysis to assess model behavior across sentiment categories. Finally, we conduct an ablation study to quantify the contribution of each architectural component. Together, these experiments offer a thorough validation of the effectiveness, robustness, and discriminative capability of the proposed multimodal architecture.

The results reported in Table 2 highlight the impact of the latent representation size on the performance of our proposed approach.

**Table 2.** Overall performance across different latent dimensions.

| Latent Dimension | Accuracy | F1-Score |
|---|---|---|
| 128 | 89.77 | 0.8978 |
| 256 | 89.31 | 0.8933 |
| 512 | 89.66 | 0.8967 |
| 1024 | 89.89 | 0.8989 |

Overall, the model demonstrates stable and competitive performance across different latent dimensions, with only slight variations in accuracy and F1-score. For example, when using a latent dimension of 128, the model achieves an accuracy of 89.77% and an F1-score of 0.8978, indicating performance even with a relatively compact representation. Increasing

the latent dimension to 256 slightly decreases performance, which may be attributed to over-parameterization or reduced generalization. As the latent dimension increases to 512, the performance improves again, suggesting that richer representations can better capture multimodal interactions. The best results are obtained with a latent dimension of 1024, achieving an accuracy of 89.89% and an F1-score of 0.8989. This improvement indicates that larger latent spaces enable the model to learn more expressive and discriminative multimodal features. Overall, these results demonstrate that our proposed approach is robust to changes in representation size while benefiting from higher-capacity latent spaces. For instance, the consistent performance across configurations confirms the effectiveness of the proposed fusion and interaction mechanisms in capturing complementary information from multiple modalities.

### *5.1. Comparison with Baselines and State-of-the-Art Methods*

Table 3 reports the performance of our proposed approach in comparison with several baseline and state-of-the-art methods on the Ar-MuSA dataset. The results demonstrate a clear and consistent improvement achieved by our model in terms of both accuracy and F1-score. For reference, the performance of MARBERT as a Text-only model on the Ar-MuSA benchmark has been previously reported in the original dataset paper and in our prior work [16,34], establishing a strong unimodal baseline for contextualizing the multimodal improvements reported in this study.

**Table 3.** Performance comparison with baselines and state-of-the-art methods in Ar-MuSA dataset.

| Method | Accuracy (%) | F1-Score |
|---|---|---|
| Late Fusion (Weighted Average) [34] | 49.0 | - |
| Late Fusion (Voting Ensemble) [34] | 63.0 | - |
| Late Fusion (SILMA AI) [17] | 68.0 | 0.65 |
| Late Fusion (LLama 3.1-8B) [17] | 71.0 | 0.67 |
| Early Fusion [16] | 77.6 | 0.776 |
| GCMA-Net [29] | 80.17 | 0.8015 |
| **Proposed approach (Ours)** | **89.89** | **0.8989** |

Late fusion strategies, including weighted averaging and voting ensembles, yield relatively limited performance, achieving 49.0% and 63.0% accuracy, respectively. More advanced late fusion approaches based on large language models, such as SILMA AI and LLama 3.1-8B, improve the results to 68.0% and 71.0% accuracy, with corresponding F1-scores of 0.65 and 0.67. Nevertheless, these methods remain constrained in their ability to model structured cross-modal interactions. Early fusion techniques further enhance performance by jointly processing multimodal representations, reaching 77.6% accuracy and an F1-score of 0.776. Similarly, GCMA-Net achieves 80.17% accuracy and an F1-score of 0.8015 by incorporating gated mechanisms for multimodal integration. In contrast, our proposed approach significantly outperforms all competing methods, attaining 89.89% accuracy and an F1-score of 0.8989. This substantial margin indicates the effectiveness of the proposed architecture in learning discriminative and well-aligned multimodal representations. Beyond quantitative improvements, a more critical comparison with recent state-of-the-art architectures reveals important methodological distinctions. For instance, early fusion approaches primarily rely on direct feature concatenation, which limits their ability to model structured cross-modal dependencies. Similarly, advanced architectures such as GCMA-Net incorporate gated cross-modal attention to enhance interaction; however, these mechanisms remain largely implicit and tend to be influenced by dominant modalities, particularly text. In contrast, the proposed approach explicitly introduces directional semantic

enhancement, enabling similarity-guided and controlled information transfer between modalities. This design allows each modality to selectively integrate complementary cues while preserving its intrinsic characteristics. Furthermore, unlike prior methods that focus on either attention-based weighting or fusion optimization, our model jointly addresses cross-modal alignment and modality imbalance through the combination of directional interaction and adaptive gating. This results in more robust and semantically coherent multimodal representations, particularly in challenging Arabic settings characterized by heterogeneous modalities and noisy signals. The observed improvement can be attributed to the integration of directional semantic enhancement and adaptive gated fusion, which jointly facilitate effective cross-modal interaction while maintaining robustness to modality-specific noise. Overall, the results confirm that our proposed approach establishes a new state-of-the-art performance on the Ar-MuSA dataset.

### *5.2. Class-Wise Performance Analysis Across Latent Space Dimensions*

To better understand the influence of the latent space dimensionality, we report the class-wise precision, recall, and F1-score for different latent dimensions, as summarized in Table 4 and illustrated in Figure 6.

**Table 4.** Class-wise performance across different latent dimensions.

| Latent Dimension | Positive | | | Negative | | | Neutral | | |
|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 |
| 128 | 0.9208 | 0.9000 | 0.9103 | 0.8812 | 0.9082 | 0.8945 | 0.8877 | 0.8901 | 0.8889 |
| 256 | 0.9345 | 0.8742 | 0.9033 | 0.8654 | 0.9184 | 0.8911 | 0.8763 | 0.8956 | 0.8859 |
| 512 | 0.9267 | 0.8968 | 0.9115 | 0.8696 | 0.9184 | 0.8933 | 0.8871 | 0.8846 | 0.8858 |
| 1024 | 0.9205 | 0.8968 | 0.9085 | 0.8829 | 0.9235 | 0.9027 | 0.8898 | 0.8874 | 0.8886 |

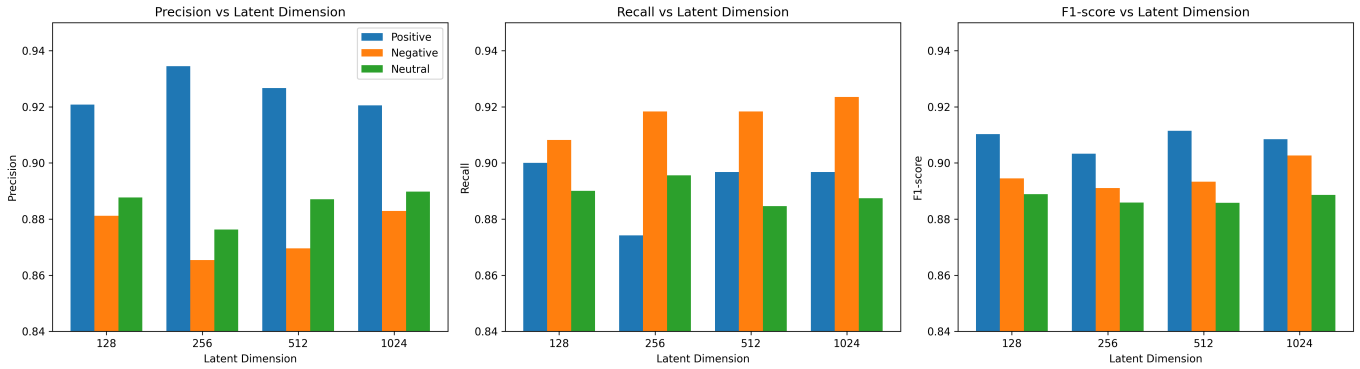



**Figure 6.** Class-wise Precision, Recall, and F1-score across different latent space dimensions (128, 256, 512, and 1024). Each subplot illustrates the performance variation for positive, negative, and neutral classes, highlighting the effect of latent space on class-specific behavior.

Overall, the results show that performance remains relatively stable across the evaluated dimensions, with only minor variations between classes. For the positive class, the precision slightly peaks at the latent dimension of 256, while recall decreases at this dimension before improving again for larger sizes. The F1-score for the positive class remains consistently high across all configurations, with the best balance observed at dimension 512. For the negative class, recall improves as the latent dimension increases, reaching its highest value at 1024, which also leads to the best F1-score for this class. However, precision for the negative class slightly decreases at intermediate dimensions before recovering at 1024, indicating a trade-off between precision and recall. In contrast, the neutral class

demonstrates the most stable behavior across all latent dimensions, with only marginal fluctuations in precision, recall, and F1-score.

The visual comparison in Figure 6 further confirms that increasing the latent dimensionality beyond 512 provides only limited improvements. While dimension 1024 slightly benefits the negative class recall, dimension 512 achieves a better balance across all classes. These findings suggest that a moderate latent dimension is sufficient to capture discriminative multimodal representations while avoiding unnecessary model complexity.

### 5.3. Ablation Study

To evaluate the contribution of each architectural component, we conduct an ablation study on the Ar-MuSA dataset. Specifically, we examine the impact of the Directional Semantic Enhancement module and the Gated Fusion mechanism by progressively removing them from the full model.

The results presented in Table 5 demonstrate the effectiveness of both the Directional Semantic Enhancement and Gated Fusion mechanisms. Removing the directional similarity component results in a noticeable decrease in performance, with accuracy dropping from 89.89% to 87.82% and F1-score declining to 0.8780. This reduction indicates that similarity-guided semantic transfer plays a critical role in modeling meaningful cross-modal dependencies. Similarly, eliminating the gated fusion mechanism leads to a performance decrease (Accuracy: 88.62%, F1: 0.8861), highlighting the importance of adaptive information control during multimodal integration. When both components are removed, the model reduces to a basic fusion strategy, yielding the lowest performance (Accuracy: 86.95%, F1: 0.8692). This cumulative degradation confirms that directional semantic enhancement and gated fusion contribute independently and synergistically to improving representation alignment and robustness. Overall, the ablation study validates that structured cross-modal interaction combined with adaptive gating is essential for achieving state-of-the-art performance in multimodal Arabic sentiment analysis.

**Table 5.** Ablation study evaluating the impact of Directional Semantic Enhancement and Gated Fusion on the Ar-MuSA dataset.

| Model Variant | Accuracy (%) | F1-Score |
|---|---|---|
| Full Model | 89.89 | 0.8989 |
| Without (DSE) [1] | 87.82 | 0.8780 |
| Without Gated Fusion | 88.62 | 0.8861 |
| Without Both | 86.95 | 0.8692 |

[1] Directional Similarity Enhancement.

It is important to note that the pretrained encoders (MARBERT, HuBERT, and ViT) provide a strong representational foundation, which explains the high baseline performance observed in the ablation study. However, the proposed DSE and Gated Fusion modules contribute an additional performance gain of approximately 3% in absolute accuracy. Although seemingly moderate, this improvement is significant in a high-performance regime and reflects enhanced cross-modal interaction, improved semantic alignment, and better handling of modality imbalance. Moreover, the proposed modules introduce limited additional parameters relative to the backbone encoders, resulting in a favorable performance–complexity trade-off.

## 6. Discussion and Implications

Multimodal Arabic sentiment analysis fundamentally relies on the effective integration of heterogeneous information sources, including text, audio, and visual signals. While

each modality conveys complementary affective cues, their structural and statistical heterogeneity introduces substantial challenges in cross-modal alignment and representation learning. The proposed directional semantic enhancement with a gated fusion approach was specifically designed to address these challenges by modeling structured cross-modal interactions while preserving modality-specific discriminative characteristics. The experimental findings validate the importance of explicitly regulating information flow between modalities rather than relying on naive concatenation or independent decision aggregation.

The results indicate that structured directional semantic transfer significantly enhances cross-modal alignment. By computing similarity-aware semantic messages between modalities and regulating their influence through learnable gating mechanisms, the model effectively reduces modality conflicts and suppresses noisy interactions. This behavior is particularly relevant in the Ar-MuSA dataset, where modality imbalance is evident, and certain modalities—especially visual frames—may contain ambiguous or weak affective signals. The directional interaction mechanism enables the model to reinforce complementary cues while preventing unreliable information from dominating the fusion space. Analysis across latent representation sizes shows that increasing fusion dimensionality boosts expressiveness. However, performance gains level off beyond moderate dimensions. This indicates that improvements are due more to structured interaction modeling than to larger representations. The architecture's control over cross-modal influence is more decisive than embedding size. This supports the idea that multimodal sentiment arises from coordinated interactions, not just feature sums. Compared to baselines and state-of-the-art methods, the proposed approach achieves better accuracy and F1-score on the Ar-MuSA benchmark. Early fusion struggles to regulate noisy cross-modal signals. Late fusion does not allow for deep semantic interaction. In contrast, the combined directional semantic enhancement and adaptive gated fusion explicitly aligns modalities while keeping the system robust. Ablation results confirm that both components contribute individually and together to performance gains.

Beyond quantitative improvements, the proposed approach carries important practical implications for real-world Arabic multimodal applications. In everyday communication, sentiment is rarely expressed solely through text; it emerges from the interplay between linguistic meaning, vocal tone, facial expressions, and contextual cues. By modeling directional cross-modal dependencies and adaptively balancing original and enhanced representations, the proposed approach offers a more realistic representation of how sentiment is conveyed in multimedia environments. Such capabilities are particularly relevant for Arabic social media video monitoring, multimedia customer feedback analysis, broadcast media sentiment tracking, and emotionally aware human–computer interaction systems. The ability to mitigate modality imbalance and handle ambiguous visual or acoustic signals strengthens the reliability of multimodal AI systems deployed in dynamic and culturally diverse Arabic contexts.

Overall, the findings underscore the importance of interaction-centric multimodal architectures. Rather than treating modalities as isolated contributors, the proposed approach demonstrates that controlled semantic transfer and adaptive gating can substantially improve representation alignment, robustness, and interpretability. These insights suggest that future multimodal Arabic sentiment research should increasingly focus on structured cross-modal reasoning mechanisms to bridge modality gaps and enhance the real-world applicability of sentiment analysis.

## 7. Conclusions

The results of this study demonstrate that the proposed Directional Semantic Enhancement with a gated fusion approach effectively improves multimodal Arabic sentiment

analysis by explicitly modeling similarity-guided cross-modal interactions. By regulating semantic transfer between text, audio, and visual modalities through adaptive gating, the architecture achieves substantial improvements over traditional early and late fusion baselines as well as recent state-of-the-art approaches. The findings confirm that structured and controlled cross-modal interaction is more effective than simple feature concatenation in mitigating modality imbalance and enhancing representation alignment. Nevertheless, several limitations remain. The evaluation is constrained to the Ar-MuSA dataset, whose size and domain coverage may limit generalization across diverse Arabic dialects and real-world multimedia environments. In particular, Ar-MuSA primarily reflects Egyptian Arabic content, and linguistic as well as acoustic characteristics vary substantially across other Arabic dialects (e.g., Gulf, Levantine, and Maghrebi varieties). Therefore, the generalization of the proposed approach to dialectally diverse settings cannot be assumed without further cross-dialect evaluation. Modality imbalance—particularly the dominance of textual cues and the ambiguity of certain visual frames—continues to pose challenges. Furthermore, although the proposed fusion mechanism improves robustness, the reliance on moderately high-dimensional latent representations increases computational complexity, which may restrict deployment in real-time or resource-constrained scenarios. Future research should investigate larger and more diverse multimodal Arabic corpora to strengthen generalization and robustness. Incorporating domain adaptation strategies may further enhance performance across varied dialectal and contextual settings. Additionally, integrating advanced pretrained multimodal transformer architectures could enrich cross-modal alignment beyond similarity-based transfer. Overall, this work underscores the promise of interaction-centric fusion mechanisms for multimodal Arabic sentiment analysis while highlighting the necessity of broader data coverage and more scalable multimodal representation learning strategies.

**Author Contributions:** Conceptualization, A.B.C.; methodology, A.B.C., E.H.N. and O.E.; software, A.B.C., E.H.N. and O.E.; validation, A.B.C., E.H.N. and O.E.; formal analysis, A.B.C., E.H.N. and O.E.; investigation, A.B.C., E.H.N. and O.E.; resources, A.B.C., E.H.N. and O.E.; data curation, A.B.C., E.H.N. and O.E.; writing—original draft preparation, A.B.C.; writing—review and editing, A.B.C., E.H.N. and O.E.; visualization, A.B.C.; supervision, E.H.N.; project administration, A.B.C. and E.H.N. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The raw data supporting the conclusions of this article will be made available by the authors on request.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. El-Idrissi, R.; Makhlouf, E.; Sahrane, M.R.; Elohounkpon, D.R.; Cristancho, J.A.S.; Reig, J.S.; Agoun, J.; Girard, J.P.; de Prado, G.; Loudcher, S.; et al. Data, Archives and Archaeological Texts: Creation and Exploitation of a Semantic Data Lake for Archaeology in Catalonia. The DataLAC Project. In Proceedings of the International Symposium Modeling the Past to Anticipate the Future, Paris, France, 4–5 December 2025.
2. Abibulaiev, A.; Pukach, P.; Vovk, M. Context-Aware ML/NLP Pipeline for Real-Time Anomaly Detection and Risk Assessment in Cloud API Traffic. *Mach. Learn. Knowl. Extr.* **2026**, *8*, 25. [CrossRef]
3. Ben Cheikhi, A.; Nfaoui, E.H. Mathematics Problem Classification Based on Pretrained Language Models. In *Proceedings of the 2025 11th International Conference on Optimization and Applications (ICOA)*; IEEE: Piscataway, NJ, USA, 2025; pp. 1–6. [CrossRef]
4. Elbiach, O.; Grissette, H.; Nfaoui, E.H. Benchmarking large language models for adverse drug reaction extraction in social media and clinical texts. *Results Eng.* **2025**, *28*, 107362. [CrossRef]

5. Ullah, M.A. BCLSA: Advancing Bangla Sentiment Analysis with Concept-Level Reasoning and Efficiency. *Mach. Learn. Appl.* **2025**, *22*, 100793. [CrossRef]
6. Chandra, R.; Ren, G. Longitudinal abuse and sentiment analysis of Hollywood movie dialogues using language models. *Mach. Learn. Appl.* **2025**, *22*, 100749. [CrossRef]
7. Mironela, P.; Iustin, P.; Mihaela, P.C.; Claudiu, P.; Daniel, G.L. Analysis of Youtube video comments with NLP methods. In *Proceedings of the 2024 16th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6.
8. Wu, H.; Kong, D.; Wang, L.; Li, D.; Zhang, J.; Han, Y. Multimodal sentiment analysis method based on image-text quantum transformer. *Neurocomputing* **2025**, *637*, 130107. [CrossRef]
9. Zhao, K.; Zheng, M.; Li, Q.; Liu, J. Multimodal sentiment analysis-a comprehensive survey from a fusion methods perspective. *IEEE Access* **2025**, *13*, 64556–64583. [CrossRef]
10. Pandey, A.; Vishwakarma, D.K. Progress, achievements, and challenges in multimodal sentiment analysis using deep learning: A survey. *Appl. Soft Comput.* **2024**, *152*, 111206.
11. Zhu, L.; Zhu, Z.; Zhang, C.; Xu, Y.; Kong, X. Multimodal sentiment analysis based on fusion methods: A survey. *Inf. Fusion* **2023**, *95*, 306–325. [CrossRef]
12. Amirthasaravanan, A.; Choudhry, M.D.; Jeevanandham, S.; Ramasamy, S.; Suvalakshmi, K.; Karthikeswaran, D. Enhanced sentiment analysis on social media using BERT and multimodal attention-based fusion. *Automatika* **2026**, *67*, 69–90.
13. Yakut, S.; Tuten, Y.T.; Caglar, E.; Aktas, M.S. A Multimodal Transformer-Based Framework for Emotion Analysis in Multilingual Video Content. *Computers* **2026**, *15*, 77. [CrossRef]
14. Haouhat, A.; Bellaouar, S.; Nehar, A.; Cherroun, H.; Abdelali, A. Arabic multimodal machine learning: Datasets, applications, approaches, and challenges. *arXiv* **2025**, arXiv:2508.12227. [CrossRef]
15. Ghorbanali, A.; Sohrabi, M.K. A comprehensive survey on deep learning-based approaches for multimodal sentiment analysis. *Artif. Intell. Rev.* **2023**, *56*, 1479–1512. [CrossRef]
16. Ben Cheikhi, A.; Nfaoui, E.H. A Transformer-Based Approach for Multimodal Arabic Sentiment Analysis. *Int. J. Adv. Comput. Sci. Appl.* **2026**, *17*, 896–906. [CrossRef]
17. Khaled, S.; Medhat, W.; Mohamed, E.H. UniTextFusion: A low-resource framework for Arabic multimodal sentiment analysis using early fusion and LoRA-tuned language models. *Ain Shams Eng. J.* **2025**, *16*, 103682.
18. Nguyen, C.D.; Nguyen, T.; Vu, D.; Tuan, L.A. Improving multimodal sentiment analysis: Supervised angular margin-based contrastive learning. In *Proceedings of the Findings of ACL: EMNLP*; Association for Computational Linguistics: Kerrville, TX, USA, 2023; pp. 14714–14724.
19. Liu, Z.; Zhou, B.; Chu, D.; Sun, Y.; Meng, L. Modality translation-based multimodal sentiment analysis under uncertain missing modalities. *Inf. Fusion* **2024**, *101*, 101973.
20. Paraskevopoulos, G.; Georgiou, E.; Potamianos, A. MMLatch: Bottom-up top-down fusion for multimodal sentiment analysis. In *Proceedings of the ICASSP*; IEEE: Piscataway, NJ, USA, 2022; pp. 4573–4577.
21. Christ, L.; Amiriparian, S.; Baird, A.; Kathan, A.; Müller, N.; Klug, S.; Gagne, C.; Tzirakis, P.; Stappen, L.; Meßner, E.M.; et al. The MuSe 2023 Multimodal Sentiment Analysis Challenge. In *Proceedings of the MuSe Workshop*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 1–10.
22. Cheikhi, A.B.; Nfaoui, E.H. Multimodal Arabic Sarcasm Detection Using CNN and BiLSTM. In *Proceedings of the International Conference on Intelligent Systems*; IEEE: Piscataway, NJ, USA, 2025; pp. 1–5.
23. Zhao, Y.; Mamat, M.; Aysa, A.; Ubul, K. Multimodal sentiment system and method based on CRNN-SVM. *Neural Comput. Appl.* **2023**, *35*, 24713–24725. [CrossRef]
24. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor Fusion Network for Multimodal Sentiment Analysis. *arXiv* **2017**, arXiv:1707.07250. [CrossRef]
25. Hori, C.; Hori, T.; Lee, T.-Y.; Zhang, Z.; Harsham, B.; Hershey, J.R.; Marks, T.K.; Sumi, K. Attention-based multimodal fusion for video description. In *Proceedings of the ICCV*; IEEE: Piscataway, NJ, USA, 2017; pp. 4193–4202.
26. Guo, Y. A mutual attention based multimodal fusion for fake news detection. *Appl. Intell.* **2023**, *53*, 15311–15320.
27. Zerkouk, M.; Mihoubi, M.; Chikhaoui, B. Contextual Attention-Based Multimodal Fusion of LLM and CNN for Sentiment Analysis. *arXiv* **2025**, arXiv:2508.13196. [CrossRef]
28. Liu, Y.; Kang, X.; Matsumoto, K.; Zhou, J. VBCSNet: A Hybrid Attention-Based Multimodal Framework with Structured Self-Attention for Sentiment Classification. *Chin. J. Inf. Fusion* **2025**, *2*, 356–369. [CrossRef]
29. Cheikhi, A.B.; Nfaoui, E.H. GCMA-Net: A Gated Cross-Modal Attention Network for Arabic Multimodal Sentiment Analysis. *IEEE Access* **2026**, *14*, 30733–30746. [CrossRef]
30. Najadat, H.; Abushaqra, F. Multimodal sentiment analysis of Arabic videos. *J. Image Graph.* **2018**, *6*, 39–43. [CrossRef]
31. Agarwal, A.; Yadav, A.; Vishwakarma, D.K. Multimodal Sentiment Analysis via RNN variants. In *Proceedings of the IEEE/ACIS International Conference on Big Data, Cloud Computing, and Data Science*; IEEE: Piscataway, NJ, USA, 2019; pp. 19–23.

32. Al-Azani, S.; El-Alfy, E.S.M. Enhanced video analytics for sentiment analysis based on fusing textual, auditory and visual information. *IEEE Access* **2020**, *8*, 136843–136857. [CrossRef]
33. Sherif, A.; Sabty, C. Sentiment analysis for Egyptian Arabic-English code-switched data. In *Proceedings of the International Conference on Speech and Computer*; IEEE: Piscataway, NJ, USA, 2024; pp. 54–69.
34. Khaled, S.; Ragab, M.E.; Helmy, A.K.; Medhat, W.; Mohamed, E.H. Ar-MuSA: A Multimodal Benchmark Dataset and Evaluation Framework for Arabic Sentiment Analysis. *Int. J. Intell. Eng. Syst.* **2025**, *18*, 30.
35. Yao, D.; Li, B.; Li, Z. Bi-directional similarity enhancement and adjustment hashing for unsupervised cross-modal retrieval. *Neurocomputing* **2026**, *672*, 132767. [CrossRef]
36. Huang, F.; Zhang, X.; Zhao, Z.; Li, Z. Bi-directional spatial-semantic attention networks for image-text matching. *IEEE Trans. Image Process.* **2018**, *28*, 2008–2020. [CrossRef]
37. Di, L.; Zhang, B.; Wang, Y.; Zhang, W. Frequency Meets Semantics: Text-Visual Fusion with Directional Spectral Enhancement for Salient Object Detection in Optical Remote Sensing Images. In *Proceedings of the 33rd ACM International Conference on Multimedia*; Association for Computing Machinery: New York, NY, USA, 2025; pp. 1987–1996.
38. Ma, Q.; Pan, J.; Bai, C. Direction-oriented visual–semantic embedding model for remote sensing image–text retrieval. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4704014. [CrossRef]
39. Zhu, L.; Wu, R.; Zhu, X.; Zhang, C.; Wu, L.; Zhang, S.; Li, X. Bi-direction label-guided semantic enhancement for cross-modal hashing. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *35*, 3983–3999. [CrossRef]
40. Guo, H.; Sun, B.; Li, S.; Lei, C.; Li, X.; Tan, M. Directional Semantic Enhanced Visual Grounding for Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4705714. [CrossRef]
41. Hendrycks, D. Gaussian Error Linear Units (Gelus). *arXiv* **2016**, arXiv:1606.08415.
42. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning. PMLR 37, Lille, France, 7–9 July 2015; pp. 448–456.
43. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
44. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2018; pp. 7132–7141.
45. Abdul-Mageed, M.; Elmadany, A.; Nagoudi, E.M.B. ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. *arXiv* **2021**, arXiv:2101.01785. [CrossRef]
46. Hsu, W.N.; Bolte, B.; Tsai, Y.H.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *arXiv* **2021**, arXiv:2106.07447.
47. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929. [CrossRef]


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.