A Transformer-Based Approach for Multimodal Arabic Sentiment Analysis

Ayoub BEN CHEIKHI, EL Habib NFAOUI
L3IA Laboratory, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco

***Abstract*—The Multimodal Sentiment Analysis (MSA) landscape for Arabic content is strikingly underexplored, mainly due to limited datasets and a lack of robust integration methods across text, audio, and image. While transformer-based models like MarBERT and ArBERT achieve strong results on Arabic text, most research remains unimodal and does not fully exploit multimodal synergy. In this work, we propose a three-fold approach for Arabic MSA. First, we finetune robust transformers for each modality, namely ViT, MarBERT, and HuBert for image, Text, and Audio, respectively. Second, we perform an early feature fusion. Third, we use classifiers for sentiment prediction. On the recent Ar-MuSA benchmark released on 2025, our tri-modal fusion system, achieves state-of-the-art performance (F1=0.7756, Accuracy=0.7759), significantly exceeding the multimodal models benchmarked on the Ar-MuSa dataset, as well as the unimodal and bimodal methods. This demonstrates that comprehensive tri-modal fusion and thoughtful classifier selection are essential for accurate, human-centric Arabic sentiment analysis.**

***Keywords*—Arabic sentiment analysis; multimodal learning; feature fusion; machine learning; early fusion**

I. Introduction

The Arabic digital space has grown rapidly in recent years [1], [2], [3], [4], [5]. Social media platforms, video-sharing services, and online forums allow users to express opinions and emotions through text, speech, and images. As a result, sentiment is increasingly communicated in a multimodal manner [6], [7], [8]. This evolution raises new challenges for artificial intelligence (AI) [9] and natural language processing (NLP) [10], which must go beyond textual analysis to understand sentiment more accurately [11]. Early work in Arabic sentiment analysis mainly focused on text written in Modern Standard Arabic (MSA) [12], [13], [14], [2]. These approaches often relied on machine learning techniques and handcrafted features [15]. The introduction of transformer-based models has significantly improved text-based sentiment analysis [16], [17]. However, Arabic remains a complex language due to its rich morphology, dialectal variation, and informal writing styles commonly used in social media. Importantly, sentiment is not expressed through text alone. Spoken language conveys emotion through prosody, such as intonation and rhythm. Visual content also plays a key role, as facial expressions and contextual cues often reveal affective states. Ignoring these signals leads to incomplete sentiment understanding, particularly in socially rich Arabic media.

While multimodal sentiment analysis has been widely explored for English [18], [6], [11], research in Arabic remains limited. One major reason is the lack of large, well-annotated multimodal datasets. Another challenge lies in designing effective fusion strategies that can integrate heterogeneous modalities. The release of the Ar-MuSA [19] dataset addresses part of this gap by providing aligned text, audio, and image data for Arabic sentiment analysis.

In this work, we investigate multimodal Arabic sentiment analysis using transformer-based feature extraction for text, speech, and images. We adopt hybrid pipelines that combine deep pretrained representations with machine learning classifiers. In particular, we focus on early feature-level fusion and systematically compare unimodal, bi-modal, and tri-modal configurations.

Beyond technical contributions, this research aims to support a more context-aware understanding of Arabic digital content. Robust multimodal sentiment analysis is essential for applications such as content moderation, opinion mining, and social media analysis. It is also especially valuable in low-resource and highly variable linguistic environments such as Arabic.

The main contributions of this study are as follows:

- We provide a systematic evaluation of multimodal sentiment analysis for Arabic using text, audio, and image modalities from the Ar-MuSA dataset.
- We analyze the impact of early fusion strategies across uni-modal, bi-modal, and tri-modal settings.
- We report the state-of-the-art results and highlight practical insights for future research in Arabic multimodal sentiment analysis.

The remainder of this study is structured as follows: Section II surveys foundational research in Arabic sentiment analysis and multimodal approaches. Section III presents the details of our proposed multimodal sentiment analysis pipeline, including unimodal feature extraction, fusion methods, and classification strategies. Section IV describes the Ar-MuSA dataset, outlines preprocessing procedures, and details the experimental protocol. Section V reports and analyzes the experimental results across unimodal, bimodal, and tri-modal configurations. Section V-B provides a direct comparison of our approach with existing Ar-MuSA benchmarks and state-of-the-art results. Section VI offers an in-depth interpretation of key findings and their broader implications. Finally, Section VII concludes the study and suggests possible directions for future research in Arabic multimodal sentiment analysis.

II. Related Work

Arabic Sentiment Analysis (ASA) has evolved from classical machine learning pipelines that rely on handcrafted features to transformer-based and multimodal approaches that integrate textual, acoustic, and visual signals. This section reviews prior

work in line with the thematic structure adopted in this study: 1) Arabic sentiment analysis, 2) multimodal Arabic sentiment analysis, 3) multimodal machine learning foundations, and 4) transformer-based classification. We conclude by synthesizing the main limitations reported in the literature and highlighting the research gap addressed by this work.

A. Arabic Sentiment Analysis

Contemporary ASA research consistently emphasizes the linguistic complexity of Arabic—including rich morphology, dialectal variation, and informal writing styles—as a central challenge for robust sentiment modeling. In [15], the authors compare conventional methods such as SVM and Naïve Bayes with deep learning and transformer-based approaches (e.g., LSTM and AraBERT), reporting that AraBERT achieves superior performance (above 88% accuracy in their reviewed settings) while noting persistent limitations related to corpus scarcity and dialectal diversity.

Several application-driven studies confirm that classical machine learning approaches remain competitive in domain-specific settings when datasets are appropriately curated. Musleh et al. investigate Arabic YouTube comments and evaluate six classifiers for binary sentiment polarity; their experiments show that Naïve Bayes achieves the best performance (94.62% accuracy and MCC of 91.46%), demonstrating the continued relevance of lightweight models for constrained ASA tasks [20]. In another social-media context, Musleh et al. propose a machine-learning approach to detect depression from Arabic Twitter posts, formulating a multi-class problem (depressed, non-depressed, neutral) and reporting that Random Forest attains 82.39% accuracy [21]. This work illustrates that extending sentiment-related analysis to mental-health signals introduces additional complexity in labeling, feature robustness, and generalization.

In the education domain, [22] construct an Arabic dataset from student satisfaction surveys and compare classical models with the AraBERT transformer, achieving an accuracy of 78% and highlighting the limited availability of Arabic sentiment datasets tailored to educational evaluation. Overall, these studies collectively show that: 1) dataset design and domain coverage are crucial to performance, and 2) pretrained transformers often provide gains, but robustness across domains and dialects remains an open challenge.

B. Multimodal Arabic Sentiment Analysis

Although text-only ASA dominates the literature, multimodal Arabic sentiment analysis (MuSA) is increasingly studied due to the prevalence of sentiment expression through voice, facial cues, and contextual imagery in videos and social platforms. Early work by [23] introduces a direction toward Arabic multimodal sentiment analysis by presenting a dataset and exploring SVM-based classification, arguing that multimodality can better capture sentiment signals in heterogeneous social media content. In [24], the authors analyze Arabic YouTube videos by integrating acoustic and facial features and testing different classifiers (including SVM and neural networks), achieving an overall accuracy of 76%, which underlines both the promise of multimodal cues and the difficulty of robust inference in real-world settings.

More recent efforts emphasize improved fusion strategies across text, audio, and vision. In [25], the authors propose enhanced video analytics through multi-level fusion of textual, auditory, and visual information, reporting more than 94% effectiveness in multi-dialect Arabic videos. Their results suggest that carefully designed fusion mechanisms can significantly improve performance beyond unimodal baselines, particularly when dialectal and expressive cues are important.

Dataset construction remains a key bottleneck for MuSA. In [26], the authors introduce and validate an Arabic multimodal dataset using transformer-based components and word-alignment techniques, but they also emphasize limitations in dataset scale and coverage. Addressing the need for standardized benchmarking, [27] propose Ar-MuSA, an open-source multimodal benchmark dataset and evaluation framework that combines text, audio, and visual elements, demonstrating that multimodal approaches often outperform unimodal strategies when validated with advanced models such as MarBERT and HuBERT. In low-resource settings, [28] further introduce UniTextFusion, an early-fusion strategy that textualizes non-text modalities to integrate them into Arabic language models; using LoRA-based fine-tuning, they report improvements of up to 34% in F1-score compared to existing baselines. These findings highlight the importance of both dataset availability and parameter-efficient multimodal fusion for Arabic.

C. Multimodal Application Using Machine Learning

General multimodal machine learning (MML) research provides the conceptual foundation for MuSA, particularly for understanding how to represent, align, and fuse heterogeneous modalities. In [29], the authors provide a comprehensive survey and propose a taxonomy organizing MML challenges into representation, translation, alignment, fusion, and co-learning. This taxonomy is especially relevant to Arabic multimodal sentiment tasks where temporal alignment and noisy cross-modal signals can substantially affect performance.

In [30], the authors extend these foundations by framing principles, challenges, and open questions in multimodal machine learning, covering representation, alignment, reasoning, generation, transference, and uncertainty quantification. Their discussion suggests that multimodal systems should be evaluated not only by accuracy but also by robustness, transferability, and reliability—criteria that are particularly important in Arabic contexts with high dialect variability and diverse media conditions.

Multimodal modeling is also increasingly applied to pragmatic phenomena closely related to sentiment, such as sarcasm. In [31], the authors present a multimodal Arabic sarcasm detection approach using text, audio, and image features on the Ar-MuSA dataset; leveraging BiLSTM, CNN, and ResNet-50 encoders, they achieve 97% accuracy and show clear gains over unimodal settings. This supports the argument that multimodality can capture implicit affective and pragmatic cues that are difficult to infer from text alone.

D. Transformer Application in Classification

Transformers have become the dominant paradigm for classification in both NLP and computer vision, and they

have strongly influenced both unimodal ASA (e.g., AraBERT-like encoders) and multimodal systems that depend on transformer backbones or transformer-compatible representations. In [32], the authors survey transformer-based text classification, proposing an expanded taxonomy and discussing performance, cost, safety, and bias considerations; importantly, they argue that large transformer models are not universally superior without careful consideration of domain and deployment constraints .

In vision, [33] review vision transformers for image classification, tracing the evolution from ViT to more advanced variants and comparing them to CNNs, while emphasizing open challenges such as data efficiency and computational demands. These insights motivate efficiency-aware multimodal designs, since MuSA systems often require strong unimodal encoders in each modality.

Evidence also indicates that smaller, well-adapted pretrained models can compete with much larger models in classification tasks. In [34], the authors evaluate pretrained language models for mathematical problem classification and show that compact transformer models can match larger LLMs with a macro-F1 of 0.8685, highlighting the practical value of targeted fine-tuning and cost-effective architectures. This aligns with recent MuSA work that favors parameter-efficient tuning methods such as LoRA [28].

E. Research Gap

Despite substantial progress, the reviewed literature reveals several persistent gaps. First, the availability of large-scale, diverse, and well-aligned Arabic multimodal datasets remains limited; while Ar-MuSA advances benchmarking [27], other dataset efforts still report constraints in scale and coverage [26]. Second, robustness across dialects, domains, and informal language remains a central challenge in ASA [15]; strong results in a specific domain (e.g., YouTube) may not transfer to other contexts such as education or mental health [20], [21], [22]. Third, although recent low-resource fusion methods show promise [28], the design space of fusion strategies (early/late/hybrid fusion and modality interaction modeling) has not been consistently evaluated under comparable settings for Arabic. Finally, many studies emphasize accuracy-focused evaluation, while broader multimodal literature argues for more systematic analysis of reliability, uncertainty, and failure modes [30].

These limitations motivate the need for an approach that: 1) better accounts for Arabic linguistic variability, 2) leverages complementary multimodal cues through principled and efficient fusion, and 3) is evaluated with robustness and generalization analysis beyond single-dataset performance.

III. Proposed Approach

The model architecture proposed in this research addresses the task of Arabic multimodal sentiment analysis through a principled integration of information from textual, visual, and audio modalities. The system leverages the strengths of transformer-based models for each modality, and employs a structured pipeline for effective fusion and classification. The process is illustrated in Fig. 1.

A. Unimodal Representation Learning

For each modality—text, audio, and image—we employ state-of-the-art transformer architectures, each fine-tuned on the Ar-MuSA dataset to extract robust, modality-specific embeddings:

$$\mathbf{e}_{\text{text}} = \text{MarBERT}(\mathbf{x}_{\text{text}}; \theta_{\text{text}}) \tag{1}$$

$$\mathbf{e}_{\text{audio}} = \text{HuBERT}(\mathbf{x}_{\text{audio}}; \theta_{\text{audio}}) \tag{2}$$

$$\mathbf{e}_{\text{image}} = \text{ViT}(\mathbf{x}_{\text{image}}; \theta_{\text{image}}) \tag{3}$$

where, $\mathbf{x}_{\text{text}}$, $\mathbf{x}_{\text{audio}}$, and $\mathbf{x}_{\text{image}}$ are the respective raw inputs, θ denotes the set of fine-tuned parameters specific to each model, and $\mathbf{e}*$ represents the resulting embedding vectors.

This approach is motivated by several key considerations. First, using dedicated transformer models for each modality allows us to capitalize on their architectural strengths, which have been empirically demonstrated across diverse tasks in natural language processing, speech analysis, and computer vision. Each model—MarBERT, HuBERT, and ViT—incorporates design choices and pretraining strategies that make it uniquely effective for its target data type.

Second, fine-tuning on the Ar-MuSA dataset ensures that each model adapts to the linguistic, acoustic, and visual characteristics specific to the domain of interest. This step enhances the models' ability to capture salient and contextually relevant patterns, which is especially important when dealing with datasets that diverge from widely available pretraining corpora.

Third, transformer-based encoders yield highly expressive and context-aware representations. These models effectively model complex dependencies: MarBERT excels at capturing nuanced semantics in Arabic text, HuBERT models both local and long-range speech patterns, and ViT efficiently encodes spatial and hierarchical information in images. The unimodal embeddings produced are therefore rich in both local detail and global structure.

Finally, a wealth of recent literature highlights the benefits of strong unimodal feature extraction as a precursor to multimodal learning [35], [36]. Isolating and optimizing each modality at the representation level not only improves single-modality interpretability, but also provides a robust foundation for effective cross-modal fusion, leading to superior performance in downstream multimodal tasks.

Through this methodical unimodal representation learning scheme, we ensure that each input stream contributes maximally informative, discriminative, and complementary features to the overall multimodal system.

B. Multimodal Feature Fusion

In multimodal learning, feature fusion refers to the process of integrating information derived from multiple input modalities—such as text, audio, and images—into a single, unified representation suitable for downstream tasks. Among

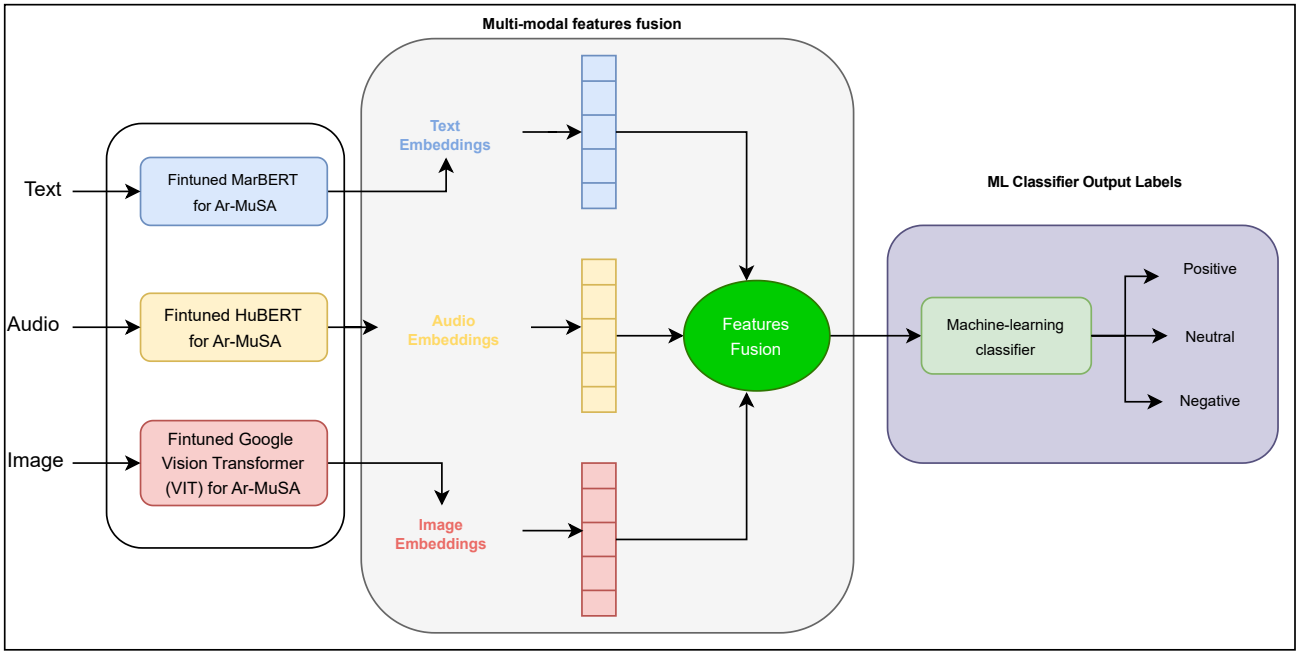

Fig. 1. Overview of the proposed multimodal Arabic sentiment analysis pipeline. Each modality is encoded by a dedicated finetuned transformer (MarBERT for text, HuBERT for audio, ViT for images). The resulting embeddings are concatenated to produce a joint multimodal feature vector, which is then used as input to a machine learning classifier (SVM) for sentiment prediction (positive, neutral, negative).

various fusion strategies, early fusion (also known as feature-level fusion) is a widely adopted approach. In early fusion, raw or low-level feature embeddings from each modality are combined at the initial stages of the model pipeline, usually before any decision-level processing or prediction.

Formally, let $\mathbf{e}_{\text{text}}$, $\mathbf{e}_{\text{audio}}$, and $\mathbf{e}_{\text{image}}$ denote the learned feature representations for the text, audio, and image modalities, respectively. Under early fusion, we concatenate these feature vectors to obtain a single, comprehensive embedding:

$$\mathbf{e}_{\text{fused}} = [\mathbf{e}_{\text{text}}; \mathbf{e}_{\text{audio}}; \mathbf{e}_{\text{image}}] \tag{4}$$

Here, $[\cdot\,;\cdot;\,\cdot]$ denotes straightforward vector concatenation.

The principal rationale for employing early fusion is its capacity to preserve the diversity and richness of information from each input modality. Unlike late fusion approaches, which combine information at the decision level and potentially disregard modality-specific patterns, early fusion allows the model to exploit nuanced interactions between modalities at a granular level. This aligns with empirical findings in recent studies, which demonstrate that feature-level fusion retains complementary signals and facilitates deeper cross-modal synergy, ultimately enhancing performance in complex multimodal learning tasks [37], [38].

C. Multimodal Sentiment Classification

Following feature extraction and fusion, the joint multimodal embedding $\mathbf{e}_{\text{fused}}$ serves as the input to a Support Vector Machine (SVM) classifier, which is tasked with sentiment prediction across three discrete categories: positive, neutral, and negative. Specifically, the SVM learns a discriminative function over the fused feature space:

$$\hat{y} = f_{\text{SVM}}(\mathbf{e}_{\text{fused}}; \mathcal{W}) \tag{5}$$

where, $\mathcal{W}$ denotes the set of weights, thresholds, and bias terms optimized during supervised training on labeled data.

The decision framework is instantiated as follows: for each training instance i, the input tuple $(\mathbf{x}^{(i)}_{\text{text}}, \mathbf{x}^{(i)}_{\text{audio}}, \mathbf{x}^{(i)}_{\text{image}})$ —processed into the fused embedding—yields a sentiment prediction:

$$(\mathbf{x}^{(i)}_{\text{text}}, \mathbf{x}^{(i)}_{\text{audio}}, \mathbf{x}^{(i)}_{\text{image}}) \rightarrow \hat{y}^{(i)} \in \{\text{positive, neutral, negative}\} \tag{6}$$

The choice of SVM for this classification task is motivated by several factors. First, SVMs are highly effective in high-dimensional settings such as those arising from feature-level fusion of transformer-based embeddings. Their emphasis on maximizing the margin between decision boundaries lends itself well to the complex, nonlinear distributions often present in multimodal data. Second, SVMs are robust to overfitting, especially in cases where the number of training samples is modest compared to the feature space dimensionality—a common scenario in multimodal sentiment datasets. The ability to use various kernel functions (e.g., linear, RBF) further enhances the SVM's flexibility to capture intricate cross-modal relationships within the fused space.

Furthermore, SVMs offer solid interpretability of the learned margins and support vectors, facilitating post-hoc anal-

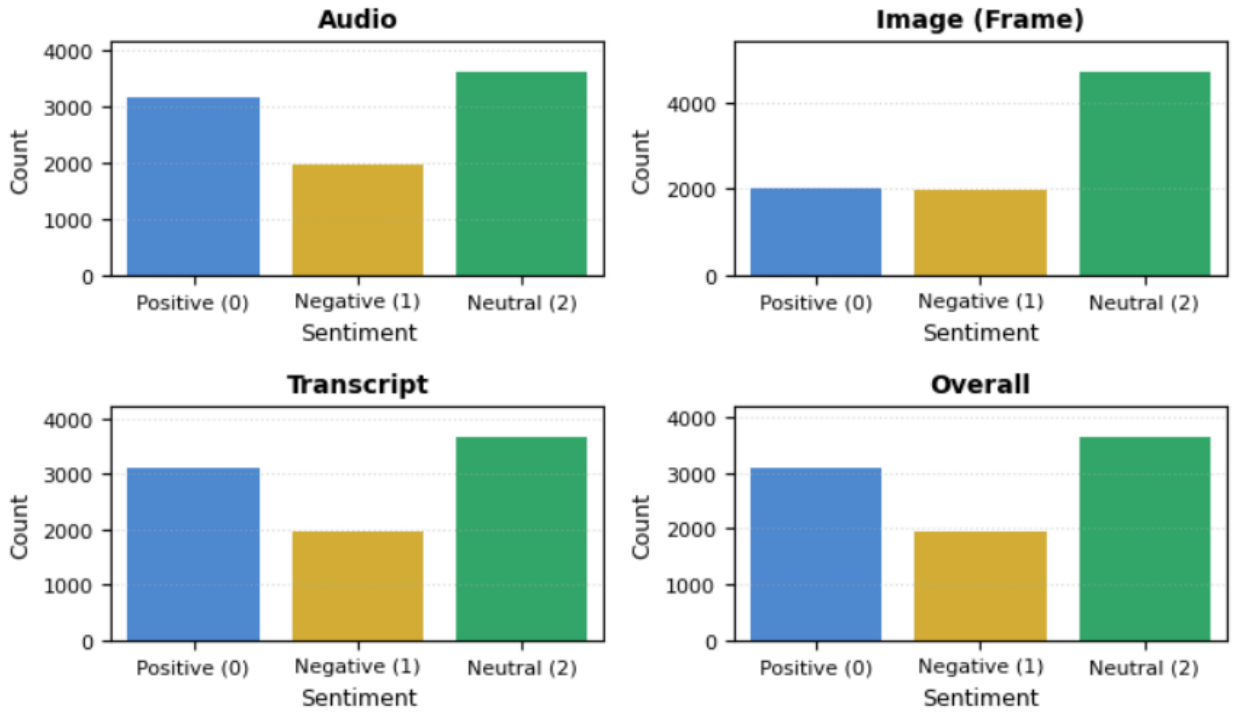

Fig. 2. Distribution of sentiment categories across modalities in the Ar-MuSA dataset.

ysis and potentially providing insights into which modalities or feature combinations most influence classification decisions. This interpretability is important not only from a research standpoint but also for transparency in consequential real-world applications such as sentiment analysis of media content or public communication.

Empirical studies corroborate the effectiveness of SVMs as final classifiers in multimodal pipelines [39], demonstrating competitive or superior performance compared to more parameter-intensive neural classifiers, especially when the preceding stages of the pipeline produce strong, information-rich feature embeddings.

In summary, the SVM-based multimodal sentiment classifier leverages the joint, complementary information captured in $\mathbf{e}_{\text{fused}}$ to robustly predict sentiment, fully exploiting the synergy of textual, acoustic, and visual modalities within a principled maximal-margin framework.

IV. Experiment

A. Dataset and Preprocessing

1) Dataset: In this study, we employ the Ar-MuSA (Arabic Multimodal Sentiment Analysis) dataset, a comprehensive and open-source benchmark specifically developed to address the scarcity of resources for Arabic multimodal sentiment analysis [19]. Ar-MuSA consists of 8,700 samples spanning three modalities—text (transcripts), audio, and images—systematically extracted from 107 Egyptian Arabic YouTube videos, with precise alignment ensured via synchronized timestamps. Each sample is annotated for sentiment (positive, negative, or neutral) at both the modality-specific and overall levels by a team of native-speaking annotators, and the final labels are determined through majority voting. The corpus significantly surpasses existing Arabic multimodal datasets in terms of both scale and diversity, providing a balanced sentiment distribution (35.5% positive, 43.3% neutral, 21.2% negative), while also capturing the unique linguistic and cultural challenges inherent in Arabic (see Fig. 2), such as sarcasm present in speech and facial expressions. The data preparation pipeline involves detailed preprocessing steps, including audio segmentation, representative frame extraction, and automated transcript generation via SeamlessM4T v2. Ar-MuSA is publicly accessible through HuggingFace, facilitating research advancements in cross-modal fusion and Arabic natural language processing.

2) Preprocessing: Preprocessing is a critical step in multimodal sentiment analysis, as it ensures consistency across modalities and prepares inputs for transformer-based modeling. In this work, we adopted modality-specific preprocessing pipelines for text, audio, and visual data to optimize compatibility with state-of-the-art models in each domain.

For the text modality, transcripts were tokenized using the MARBERT tokenizer, which is designed explicitly for Arabic, incorporating both dialectal variants and Modern Standard Arabic [40]. During preprocessing, raw text was mapped to token indices, and each sequence was padded or truncated to a maximum length of 128 tokens to maintain uniformity across samples. This standardization ensures consistent input dimensions for the MARBERT model during training and inference.

The audio modality comprised segmented utterances aligned with the text and visual data. Audio signals were first converted to mono and resampled to a sampling rate of 16 kHz. Following this, we utilized HuggingFace's

Wav2Vec2FeatureExtractor for feature extraction, which included normalization, framing, and padding or truncation to a fixed duration of 5 seconds (approximately 80,000 samples). This approach ensures that each audio sequence is of comparable length, facilitating efficient batch processing and stable learning during model fine-tuning.

We extracted a single frame per utterance corresponding to the relevant audio and text segments for the visual modality. Each image was loaded in RGB format to ensure color consistency across the dataset. Preprocessing was performed using the ViTImageProcessor from the Hugging Face Transformers library. This extractor standardized all frames to the required 224 × 224-pixel size and applied intensity normalization, preparing the images for input into the Vision Transformer model ("google/vit-base-patch16-224")[41].

B. Training Details

In this work, our primary focus is maximizing the expressive power of tri-modal fusion for Arabic multimodal sentiment analysis. While we report baseline results for unimodal and bimodal scenarios for comparative rigor, most of our methodological attention is dedicated to the optimal training and integration of the textual, acoustic, and visual streams in the tri-modal setting. To this end, our training protocol is structured around two core stages: 1) independent modality-specific fine-tuning and 2) downstream fusion and classifier training. Below, we provide a detailed account of both, explicitly referencing all training protocols and model configurations utilized.

The specific hyperparameters for each modality were determined empirically and are summarized in Table I, Table II, and Table III. These tables offer a transparent and comprehensive reference with all key settings, including optimizer type (Adam or AdamW), learning rate, batch size, number of epochs, gradient accumulation steps, warmup steps or ratios, evaluation frequency, and initialization seeds.

TABLE I. TRAINING SETUP FOR MARBERT MODEL

Parameter	Value
Adam Epsilon	1e-8
Learning Rate	1e-5
FP16	Enabled
Train Batch Size	16
Eval Batch Size	16
Gradient Accumulation Steps	2
Number of Epochs	5
Warmup Steps	0.2
Evaluation Strategy	Per Epoch
Random Seed	42
Logging Steps	200

After feature extraction from the fine-tuned MarBERT (text), HuBERT (audio), and ViT (visual) models, the respective feature vectors were concatenated along the feature dimension to form unified representations for each data instance. These tri-modal concatenated feature vectors were then used as input for a suite of machine learning classifiers: Support Vector Machine (SVM), Random Forest (RF), k-nearest Neighbors (KNN), and Naive Bayes (NB), and each

TABLE II. TRAINING SETUP FOR HUBERT MODEL

Parameter	Value
Adam Epsilon	1e-8
Learning Rate	2e-5
FP16	Enabled
Train Batch Size	8
Eval Batch Size	8
Gradient Accumulation Steps	2
Number of Epochs	3
Warmup Steps	0.2
Evaluation Strategy	Per Epoch
Random Seed	42
Logging Steps	200

TABLE III. TRAINING SETUP FOR VIT 16X16 MODEL

Parameter	Value
Train Batch Size	32
Eval Batch Size	32
Learning Rate	0.0001
Input_shape	(224, 224, 3)
Optimizer	Adam
Loss_function	Categorical Cross entropy
Number of Epochs	3
Metrics	Accuracy

model was implemented within a `scikit-learn` pipeline [42]. Table IV summarizes the primary hyperparameters used for each classifier, as optimized or specified for our tri-modal (Text+Audio+Image) classification experiments:

For all experiments, data were partitioned using a stratified split, with 80% of the data allocated for training, 10% for validation, and 10% for testing, while maintaining label distribution across all subsets (`random_state=42`) to ensure reproducibility. Unless specified otherwise, all classifier hyperparameters were tuned and evaluated primarily on the tri-modal feature fusion setting, which constitutes the central focus of our approach.

C. Evaluation Metrics

A rigorous evaluation of our multimodal sentiment classification model requires a multifaceted approach, particularly given the frequent class imbalance typical of sentiment analysis datasets. To this end, we adopted a comprehensive suite of metrics to ensure fairness and depth in assessing our models. The primary metrics considered include overall accuracy, weighted precision (WP), weighted recall (WR), and weighted F1-score (WF-1). Below, we present explicit definitions and the mathematical formulations employed for each metric.

1) Weighted Precision (WP): Precision quantifies the proportion of optimistic predictions that are indeed correct. In scenarios where classes are imbalanced, simply averaging across classes can be misleading. Weighted precision mitigates this by giving higher importance to classes with greater representation in the data. It is computed as:

$$\text{WP} = \frac{\sum_{i=1}^{n} [\text{Precision}_i \times \text{Support}i]}{\sum i = 1^n \text{Support}_i}$$

where, Precision_i is the precision for class i, Support_i is

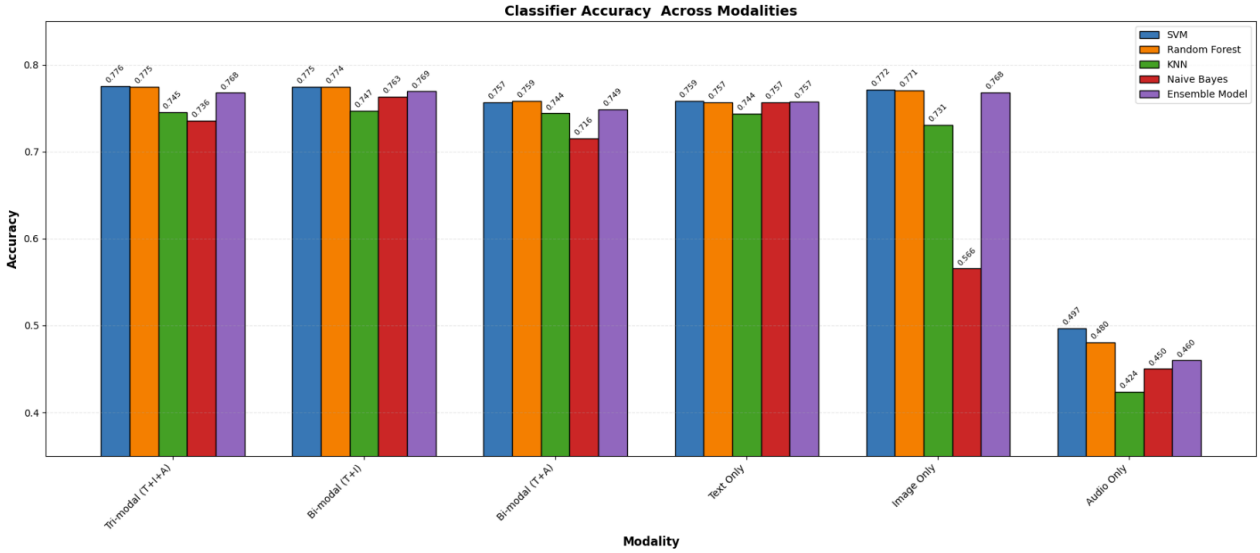

Fig. 3. Classification accuracy (%) of all evaluated classifiers across modalities and fusion types.

TABLE IV. PRIMARY HYPERPARAMETERS USED FOR EACH CLASSIFIER IN THE TRI-MODAL EXPERIMENT

Classifier	Hyperparameters
SVM (RBF)	kernel=rbf C=1.0 gamma=scale probability=True random_state=42
Random Forest	n_estimators=500 max_depth=30 max_features=sqrt class_weight=balanced min_samples_split=5 n_jobs=-1 random_state=42
KNN	n_neighbors=5 weights=distance algorithm=auto n_jobs=-1
Naive Bayes	GaussianNB après normalisation)
Ensemble (Voting)	Voting=soft estimateurs={SVM, RF, KNN, NB}

the number of true samples in class i, and n is the total number of classes.

2) Weighted Recall (WR): Recall measures the ability of the classifier to identify all relevant instances for each class correctly. Just like precision, it is important to adjust for class support. Weighted recall is defined as:

$$\text{WR} = \frac{\sum_{i=1}^{n} [\text{Recall}_i \times \text{Support}_i]}{\sum_{i=1}^{n} \text{Support}_i}$$

with Recall_i representing the recall for class i. This metric ensures a more representative assessment, especially when certain classes dominate the dataset.

3) Weighted F1-Score (WF-1): The F1-score is the harmonic mean of precision and recall and provides a single measure of a classifier's accuracy about the positive class. The weighted F1-score, which addresses the variance in class sizes, is calculated as:

$$\text{WF-1} = \frac{\sum_{i=1}^{n} [\text{F1-Score}_i \times \text{Support}_i]}{\sum_{i=1}^{n} \text{Support}_i}$$

where, the F1-score for each class, i, is given by:

$$\text{F1-Score}_i = \frac{2 \times \text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$$

4) Accuracy: Often considered the most intuitive metric, accuracy is the ratio between the number of correct predictions and the total number of predictions made:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

To provide additional transparency, each experimental result is accompanied by a full classification report detailing precision, recall, F1-score, and support for each sentiment class (positive, negative, and neutral). We further visualize performance using confusion matrices, which offer insight into which classes the model will most likely confuse.

All metrics are systematically computed using the `scikit-learn` Python package [42], ensuring reproducibility and consistency across experiments. Using weighted metrics throughout is deliberate, as it more accurately reflects classifier performance in the presence of uneven class distributions—an everyday reality in real-world sentiment datasets.

This robust evaluation protocol allows for a nuanced comparison of different classifiers and fusion methods, underpinning the credibility of our findings in multimodal sentiment analysis.

TABLE V. COMPARISON OF OUR CONFIGURATIONS WITH STATE-OF-THE-ART SYSTEMS AND AR-MUSA BENCHMARK BASELINES

System / Study	Modality / Fusion	Model	WF-1	Accuracy
Our (Tri-modal)	Text + Audio + Image	SVM	**0.776**	**0.776**
Khaled et al. [28]	Late Fusion	LLaMA 3.1–8B	0.67	0.71
Khaled et al. [28]	Late Fusion	SILMA AI	0.65	0.68
Ar-MuSA [19]	Late Fusion	Voting Ensemble	0.60	0.63

V. RESULTS

This section presents an in-depth evaluation of our multimodal sentiment classification methodology on the Ar-MuSA benchmark, particularly emphasizing the tri-modal (text, audio, and image) fusion setting. For broader benchmarking, unimodal and bi-modal results are included. However, the analysis prioritizes tri-modal fusion—demonstrating its integrative strengths and practical relevance for robust Arabic sentiment recognition, as shown in Fig. 3.

A. Main Focus: Tri-modal Fusion Results

Integrating textual, audio, and visual streams is hypothesized to leverage complementary evidence and improve sentiment classification performance. In our tri-modal fusion, we concatenate deep feature vectors from MarBERT (text), HuBERT (audio), and ViT (image), yielding a high-dimensional representation for each instance. Table VI documents performance across classifiers.

TABLE VI. TRI-MODAL SENTIMENT CLASSIFICATION RESULTS (TEXT–IMAGE–AUDIO FUSION)

Classifier	Accuracy	WR	WF-1	WP
SVM	**0.7759**	**0.7759**	**0.7756**	**0.7783**
Random Forest	0.7747	0.7747	0.7747	0.7777
KNN	0.7454	0.7454	0.7449	0.7466
Naive Bayes	0.7356	0.7356	0.7362	0.7386
Ensemble Model	0.7678	0.7678	0.7678	0.7691

Tri-modal fusion consistently yields the highest overall performance, confirming the value of aggregating diverse cues. The SVM outperforms all other models (Accuracy: 0.7759, WF-1: 0.7756), suggesting that synergistic evidence from all three modalities enables stronger discrimination of sentiment—especially capturing aspects that cannot be derived from text or image alone. Random Forest is a close second, while KNN and Naive Bayes lag, possibly due to their limitations with complex, high-dimensional data. The ensemble, while robust, does not exceed the single SVM, indicating that hard or soft voting does not add to the already dominant SVM classification when features are so robust. Importantly, tri-modal fusion provides measurable, though sometimes modest, gains compared to unimodal and bimodal systems, underscoring its unique value.

B. Comparison with the State-of-the-Art and Baselines

To quantify the effectiveness of our approach, we compare our best modal configurations against the official Ar-MuSA benchmark baselines [19], as well as the recent late-fusion large-model results reported by Khaled et al. [28]. Following common practice in sentiment classification on potentially imbalanced datasets, we report weighted F1-score (WF-1) and accuracy as primary metrics. Table V provides a consolidated view of the competing systems and highlights the relative contribution of each modality and fusion setting in our pipeline. For an intuitive overview of the relative gains across systems.

As observed in Table V, our tri-modal configuration (Text+Audio+Image) with early feature concatenation and an SVM classifier achieves a WF-1 of 0.776 and an accuracy of 0.776, establishing the overall results among the compared systems. In particular, this represents a clear improvement over the best tri-modal baseline reported in Ar-MuSA (late-fusion voting ensemble; WF-1 = 0.60, accuracy = 0.63) [19], indicating that our fusion and feature design enables more effective exploitation of complementary cross-modal sentiment cues. Compared with recent late-fusion, large-model approaches [28], our best configurations remain consistently both metrics. This outcome highlights that, for Ar-MuSA-style sentiment analysis, carefully designed early fusion coupled with effective feature extraction can rival (and in this case outperform) heavier late-fusion strategies, while remaining comparatively simple and potentially more efficient to train and deploy.

C. Uni and Bi-modal Results: Added for Comparative Benchmarking

To contextualize the strength of tri-modal fusion, we evaluated unimodal and bimodal setups—especially for comparison with the Ar-MuSA baseline. These results reveal how much each modality contributes individually and in pairs, but serve primarily as references for the added value of tri-modal integration.

1) Unimodal results: Table VII compiles unimodal results across text, audio, and image—for all classifiers.

TABLE VII. UNIMODAL SENTIMENT CLASSIFICATION RESULTS ACROSS MODALITIES.

Modality	Classifier	Accuracy	WR	WF-1	WP
Text	SVM	**0.7586**	**0.7586**	**0.7580**	**0.7595**
	Random Forest	0.7569	0.7569	0.7563	0.7584
	KNN	0.7437	0.7437	0.7433	0.7436
	Naive Bayes	0.7569	0.7569	0.7574	**0.7595**
	Ensemble Model	0.7575	0.7575	0.7577	0.7587
Audio	SVM	**0.4971**	**0.4971**	**0.4602**	**0.4887**
	Random Forest	0.4805	0.4805	0.4549	0.4588
	KNN	0.4236	0.4236	0.4180	0.4151
	Naive Bayes	0.4500	0.4500	0.4537	0.4772
	Ensemble Model	0.4598	0.4598	0.4632	0.4775
Image	SVM	**0.7718**	**0.7718**	**0.7641**	**0.7687**
	Random Forest	0.7708	0.7708	0.7631	0.767 7
	KNN	0.7310	0.7310	0.7263	0.7259
	Naive Bayes	0.5655	0.5655	0.5746	0.6871
	Ensemble Model	0.7684	0.7684	0.7700	0.7727

Text and image provide height unimodal baselines, but each is limited: image achieves the highest scores (SVM WF-1: 0.7641) yet cannot capture language-specific sentiment; text is robust (SVM WF-1: 0.758) but may miss paralinguistic or non-verbal signals. Audio is the weakest (SVM WF-1: 0.460), reflecting the challenge of sentiment inference from prosody alone and possible imbalance in audio cues, a point corroborated by prior results on Arabic datasets. Ultimately, these unimodal numbers highlight why fusion—especially tri-modal fusion—is necessary for holistic sentiment recognition.

2) Bi-modal results: Bi-modal fusion results (see Table VIII) are listed next. Here, we focus on the text-audio and text-image pairs, following the Ar-MuSA evaluation.

TABLE VIII. BI-MODAL SENTIMENT CLASSIFICATION RESULTS (TEXT-AUDIO AND TEXT-IMAGE FUSION).

Fusion	Classifier	Accuracy	WR	WF-1	WP
T+A	SVM	0.7569	0.7569	0.7562	0.7583
	Random Forest	**0.7586**	**0.7586**	**0.7581**	**0.7602**
	KNN	0.7443	0.7443	0.7439	0.7441
	Naive Bayes	0.7155	0.7155	0.7158	0.7165
	Ensemble Model	0.7489	0.7489	0.7487	0.7493
T+I	**SVM**	**0.7750**	**0.7750**	**0.7746**	**0.7776**
	Random Forest	0.7745	0.7745	0.7743	0.7774
	KNN	0.7471	0.7471	0.7467	0.7482
	Naive Bayes	0.7632	0.7632	0.7637	0.7664
	Ensemble Model	0.7695	0.7695	0.7697	0.7716

Notes: T = Text modality; A = Audio modality; I = Image modality.

Bi-modal fusion, particularly with the text-image pair, yields performance only slightly above the best unimodal results (WF-1 up to 0.7746 for SVM). In some cases, improvements are marginal or classifier-dependent. The text-audio pair offers minimal additional information over text alone, which aligns with the weaker standalone audio performance. These outcomes reinforce that while bimodal fusion can harness some complementary signals, the dominant contribution still arises from the text stream, and these combinations primarily serve as strong baselines and ablation checks for the tri-modal configuration.

It is evident from the above results that while unimodal and bi-modal settings offer competitive baselines and are crucial for benchmarking against prior work, such as the Ar-MuSA dataset, their practical ceiling is reached quickly. The tri-modal approach remains unmatched in providing robust and nuanced sentiment analysis by integrating all available cues. These findings support the growing consensus in the literature that multimodal fusion—when effectively implemented—can yield modest but consistent advances over single or dual-modality systems, especially in linguistically and emotionally rich domains such as Arabic social media.

In the subsequent section, we contextualize these findings by directly comparing our results to the Ar-MuSA benchmarks, highlighting the gains achieved with our tri-modal fusion strategy.

VI. DISCUSSION

The experimental results clearly demonstrate the advantages of multimodal learning for Arabic sentiment analysis on social media data. Across all evaluated settings, tri-modal fusion (Text+Audio+Image) consistently achieved the best performance, with early feature concatenation combined with an SVM classifier yielding the highest scores (WF-1 = 0.776, Accuracy = 0.776). This confirms that sentiment in user-generated Arabic content is not solely conveyed through text, but rather emerges from the complementary interaction of linguistic, visual, and acoustic cues.

Unimodal experiments highlight the relative contribution of each modality. Text and image independently provide strong sentiment signals, whereas audio alone remains comparatively weak due to noise, speaker variability, and the subtle nature of prosodic cues. Nevertheless, when integrated with text and image, audio contributes complementary information that improves robustness and reduces ambiguity in difficult cases.

Bi-modal fusion experiments reveal that Text+Image represents the most effective dual-modality configuration, achieving performance close to the full tri-modal system. In contrast, Text+Audio fusion offers only marginal gains over text-only models. These findings suggest that visual context plays a particularly important role in disambiguating sentiment in Arabic social media, while audio acts as a secondary but supportive modality.

Compared with prior work and official Ar-MuSA baselines, the proposed approach consistently outperforms existing unimodal, bi-modal, and tri-modal systems, including recent late-fusion and large-model approaches. Notably, the performance of a relatively simple early-fusion SVM pipeline indicates that carefully designed feature-level integration of pretrained encoders can be more effective than heavier ensemble or late-fusion architectures.

Overall, the results advocate for comprehensive tri-modal fusion as the most reliable strategy for sentiment classification in Arabic social media, while also highlighting Text+Image fusion as a practical near-optimal alternative when audio is unavailable. These findings reinforce the importance of multimodal representations for capturing the richness and variability of sentiment expression in low-resource and highly informal linguistic settings.

VII. CONCLUSION

This study introduced a comprehensive framework for Arabic multimodal sentiment analysis that integrates textual, visual, and acoustic information through early feature-level fusion. By leveraging strong pretrained encoders—MarBERT for text, ViT for images, and HuBERT for speech—and combining them with margin-based classification, the proposed approach achieves state-of-the-art performance on the Ar-MuSA benchmark. Experimental results confirm that tri-modal fusion consistently outperforms unimodal and bi-modal configurations, underscoring the importance of jointly modeling linguistic content, visual context, and paralinguistic cues for sentiment understanding in user-generated Arabic media.

The study further demonstrates that while text and image modalities provide strong standalone signals, audio plays a complementary role that enhances robustness when integrated with other modalities, despite its relatively weaker individual performance. Compared with existing baselines and recent

late-fusion approaches, the proposed early-fusion strategy offers a more effective and computationally efficient alternative.

Building on these findings, future work will focus on extending the framework through improved speech representation learning, dialect-aware acoustic modeling, and more interaction-aware fusion strategies. Such extensions aim to further strengthen multimodal representations and advance sentiment analysis in low-resource, highly variable Arabic social media environments.

REFERENCES

[1] S. Gadomski, "North african arabic literary expressions of lgbtqia+ identity in the digital space," in *Digital Expressions of Gender in Africa*. Routledge, 2025, pp. 50–70.

[2] A. Habberrih and M. A. Abuzaraida, "Sentiment analysis of arabic dialects: A review study," in *International Conference on Computing and Informatics*. Springer, 2023, pp. 137–153.

[3] I. Hadjadji, L. Falek, L. Demri, and H. Teffahi, "Emotion recognition in arabic speech," in *2019 international conference on advanced electrical engineering (ICAEE)*. IEEE, 2019, pp. 1–5.

[4] N. Boudad, R. Faizi, R. O. H. Thami, and R. Chiheb, "Sentiment analysis in arabic: A review of the literature," *Ain Shams Engineering Journal*, vol. 9, no. 4, pp. 2479–2490, 2018.

[5] S. Klaylat, Z. Osman, L. Hamandi, and R. Zantout, "Emotion recognition in arabic speech," *Analog Integrated Circuits and Signal Processing*, vol. 96, no. 2, pp. 337–351, 2018.

[6] R. Das and T. D. Singh, "Multimodal sentiment analysis: a survey of methods, trends, and challenges," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–38, 2023.

[7] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," *Information Fusion*, vol. 91, pp. 424–444, 2023.

[8] T. Zhu, L. Li, J. Yang, S. Zhao, H. Liu, and J. Qian, "Multimodal sentiment analysis with image-text interaction network," *IEEE transactions on multimedia*, vol. 25, pp. 3375–3385, 2022.

[9] P. H. Winston, *Artificial intelligence*. Addison-Wesley Longman Publishing Co., Inc., 1984.

[10] J. O'Connor and I. McDermott, *NLP*. Thorsons London, UK:, 2001.

[11] S. Lai, X. Hu, H. Xu, Z. Ren, and Z. Liu, "Multimodal sentiment analysis: A survey," *Displays*, vol. 80, p. 102563, 2023.

[12] S. Alfarhood, "Soutcom: Real-time sentiment analysis of arabic text for football fan satisfaction using a bidirectional lstm," *Expert Systems*, vol. 42, no. 2, p. e13641, 2025.

[13] E. H. Nfaoui and H. Elfaik, "Evaluating arabic emotion recognition task using chatgpt models: a comparative analysis between emotional stimuli prompt, fine-tuning, and in-context learning," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 19, no. 2, pp. 1118–1141, 2024.

[14] S. E. Alqaan and A. M. Qamar, "Sentiment analysis of arabic tweets on online learning during the covid-19 pandemic: A machine learning and lstm approach," *Ingenierie des Systemes d'Information*, vol. 28, no. 6, p. 1435, 2023.

[15] S. Firdous and M. S. Iqbal, "Exploring contemporary arabic sentiment analysis: Methods, challenges, and future trends," *Pakistan Journal of Multidisciplinary Innovation*, vol. 4, no. 1, pp. 34–48, 2025.

[16] R. Zarnoufi, "Maproc at ahasis shared task: Few-shot and sentence transformer for sentiment analysis of arabic hotel reviews," *arXiv preprint arXiv:2511.15291*, 2025.

[17] A. Albladi, M. Islam, and C. Seals, "Sentiment analysis of twitter data using nlp models: a comprehensive review," *IEEE Access*, 2025.

[18] H. Yang, Y. Zhao, Y. Wu, S. Wang, T. Zheng, H. Zhang, Z. Ma, W. Che, S. Wang, S. Wei *et al.*, "Large language models meet text-centric multimodal sentiment analysis: A survey," *Science China Information Sciences*, vol. 68, no. 10, pp. 1–29, 2025.

[19] 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 & Systems*, vol. 18, no. 4, 2025.

[20] D. A. Musleh, I. Alkhwaja, A. Alkhwaja, M. Alghamdi, H. Abahussain, F. Alfawaz, N. Min-Allah, and M. M. Abdulqader, "Arabic sentiment analysis of youtube comments: Nlp-based machine learning approaches for content evaluation," *Big Data and Cognitive Computing*, vol. 7, no. 3, p. 127, 2023.

[21] D. A. Musleh, T. A. Alkhales, R. A. Almakki, S. E. Alnajim, S. K. Almarshad, R. S. Alhasaniah, S. S. Aljameel, and A. A. Almuqhim, "Twitter arabic sentiment analysis to detect depression using machine learning," *Computers, Materials & Continua*, vol. 71, no. 2, 2022.

[22] H. Alamoudi, N. Aljojo, A. Munshi, A. Alghoson, A. Banjar, A. Tashkandi, A. Al-Tirawi, and I. Alsaleh, "Arabic sentiment analysis for student evaluation using machine learning and the arabert transformer," *Engineering, Technology & Applied Science Research*, vol. 13, no. 5, pp. 11 945–11 952, 2023.

[23] 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.

[24] H. Najadat and F. Abushaqra, "Multimodal sentiment analysis of arabic videos," *Journal of Image and Graphics*, vol. 6, no. 1, pp. 39–43, 2018.

[25] 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.

[26] A. Haouhat, S. Bellaouar, A. Nehar, and H. Cherroun, "Towards arabic multimodal dataset for sentiment analysis," in *2023 Fourth International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*. IEEE, 2023, pp. 126–133.

[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 & Systems*, vol. 18, no. 4, 2025.

[28] 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.

[29] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[30] P. P. Liang, A. Zadeh, and L.-P. Morency, "Foundations & trends in multimodal machine learning: Principles, challenges, and open questions," *ACM Computing Surveys*, vol. 56, no. 10, pp. 1–42, 2024.

[31] A. B. Cheikhi and E. H. Nfaoui, "Multimodal arabic sarcasm detection using cnn and bilstm," in *2025 International Conference on Intelligent Systems: Theories and Applications (SITA)*, 2025, pp. 1–5.

[32] J. Fields, K. Chovanec, and P. Madiraju, "A survey of text classification with transformers: How wide? how large? how long? how accurate? how expensive? how safe?" *IEEE Access*, vol. 12, pp. 6518–6531, 2024.

[33] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, and L. Wang, "Vision transformers for image classification: A comparative survey," *Technologies*, vol. 13, no. 1, p. 32, 2025.

[34] A. Ben Cheikhi and E. H. Nfaoui, "Mathematics problem classification based on pretrained language models," in *2025 11th International Conference on Optimization and Applications (ICOA)*, 2025, pp. 1–6.

[35] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, and Y. Zong, "A survey of deep learning-based multimodal emotion recognition: Speech, text, and face," *Entropy*, vol. 25, no. 10, p. 1440, 2023.

[36] A. A. Gladys and V. Vetriselvi, "Survey on multimodal approaches to emotion recognition," *Neurocomputing*, vol. 556, p. 126693, 2023.

[37] H. Elfaik *et al.*, "Leveraging feature-level fusion representations and attentional bidirectional rnn-cnn deep models for arabic affect analysis on twitter," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 1, pp. 462–482, 2023.

[38] B. Azahouani, H. Elfaik, S. El Garouani *et al.*, "Multimodal sarcasm detection method using rnn and cnn," in *2024 Sixth International Conference on Intelligent Computing in Data Sciences (ICDS)*. IEEE, 2024, pp. 1–6.

[39] O. Alharbi, "A deep learning approach combining cnn and bi-lstm with svm classifier for arabic sentiment analysis," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, 2021.

[40] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "Arbert & marbert: Deep bidirectional transformers for arabic," 2021. [Online]. Available: https://arxiv.org/abs/2101.01785

[41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.