

MSAD-Moroccan dialect: a multimodal sentiment analysis dataset for Moroccan dialect (Darija)

Ayoub Ben Cheikhi, El Habib Nfaoui

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

Article Info

Article history:

Received Jan 20, 2026

Revised Jun 27, 2026

Accepted Jul 18, 2026

Keywords:

Feature fusion

Machine learning

Moroccan dialect (Darija)

Multimodal dataset

Multimodal sentiment analysis

ABSTRACT

Sentiment analysis for Arabic dialects remains a significant challenge, mainly due to the scarcity of annotated resources and the inherently multimodal nature of real-world human communication. Among Arabic dialects, Moroccan Darija is particularly underrepresented in the literature, as most existing datasets focus exclusively on textual modality. To bridge this gap, we present a multimodal sentiment analysis dataset (MSAD)–Moroccan dialect, a multimodal dataset built from authentic Moroccan podcast videos, integrating synchronized textual, acoustic, and visual representations annotated across three sentiment classes: positive, negative, and neutral. To evaluate the dataset, we conducted baseline experiments using unimodal, bimodal, and trimodal configurations with support vector machines (SVM) and multi-layer perceptrons (MLP) classifiers under various fusion strategies. Our results consistently show that multimodal configurations outperform their unimodal counterparts. Notably, the trimodal setup achieved the best overall performance, reaching an accuracy of 0.6176 and a weighted F1-score of 0.6019 using an MLP classifier with concatenation-based fusion. These findings highlight the effectiveness of multimodal learning for sentiment understanding in conversational contexts, and position MSAD as a new benchmark for sentiment analysis in low-resource Arabic dialects.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Ayoub Ben Cheikhi

L3IA Laboratory, Department of Computer Science, Faculty of sciences Dhar El Mahraz

Sidi Mohamed Ben Abdellah University

Fez, Morocco

Email: ayoub.bencheikhi@usmba.ac.ma

1. INTRODUCTION

Multimodal language understanding has become an important research direction in artificial intelligence because human communication naturally involves multiple complementary signals such as text, speech, and visual cues. By jointly analyzing these modalities, computational models can capture richer contextual information and improve the interpretation of human emotions, opinions, and intentions.

Recent research has highlighted the importance of multimodal representation learning and information fusion for modeling complex communication signals. Several studies have explored how integrating heterogeneous modalities improves the performance of machine learning systems across various applications such as recognition systems, multimedia analysis, and human–computer interaction [1]–[3]. Comprehensive surveys have also demonstrated that multimodal deep learning architectures allow models to leverage complementary signals from different modalities, leading to improved performance in tasks such as sentiment analysis and

emotion recognition [4]. Additional works have proposed multimodal frameworks combining audio, textual, and visual data for complex decision-making problems and recognition systems [5], [6]. In contrast, traditional sentiment analysis approaches have primarily relied on textual data. Early research on sentiment analysis demonstrated that linguistic features extracted from online text can be used to automatically identify opinion polarity [7], [8]. Deep learning models such as recurrent neural networks and bidirectional long short-term memory (BiLSTM) architectures further improved sentiment classification by capturing contextual relationships within textual sequences [9]. A comprehensive survey of sentiment analysis methods also highlights the dominance of text-based techniques and discusses their applications and challenges [10].

However, relying exclusively on textual information can overlook important sentimental cues conveyed through speech prosody, facial expressions, and visual context. For this reason, multimodal sentiment analysis has received increasing attention in recent years. Several surveys emphasize that combining textual, acoustic, and visual modalities can significantly improve the detection of affective states and better reflect the richness of real-world communication [11]–[13]. These studies also highlight the importance of multimodal datasets for enabling the development and evaluation of robust sentiment analysis systems.

In our recent work, we introduced a gated cross-modal attention network designed to address modality imbalance and weak cross-modal alignment in Arabic multimodal sentiment analysis [14]. The proposed architecture integrates modality projection layers, cross-modal attention mechanisms, and a reliability-aware gated fusion strategy that adaptively balances multimodal features. Experimental results demonstrate that this approach significantly improves performance over traditional fusion strategies and recent baselines, achieving an accuracy of 80.17% and an F1-score of 0.8015. Despite these advances, research on multimodal sentiment analysis for Arabic remains limited. Existing datasets mainly focus on high-resource languages or rely exclusively on textual data derived from social media platforms [15], [16]. Moroccan dialect (Darija), in particular, presents unique challenges due to its linguistic divergence from Modern Standard Arabic, frequent code-switching, and reliance on spoken communication.

To overcome existing limitations, this paper introduces multimodal sentiment analysis dataset (MSAD)-Moroccan dialect, a novel MSAD constructed from authentic Moroccan podcast videos. The dataset provides synchronized textual, acoustic, and visual representations extracted from natural conversational segments and manually annotated by native speakers. Figure 1 illustrates an example sample. By integrating these aligned modalities, MSAD offers a realistic benchmark for modeling sentiment expression in spoken Moroccan dialect. Beyond its linguistic contribution, MSAD also supports research at the intersection of artificial intelligence and hardware-oriented systems. As multimodal sentiment analysis requires processing heterogeneous data streams, the dataset enables the development of optimized multimodal pipelines suitable for deployment on embedded, reconfigurable, and very large scale integration (VLSI)-based platforms. To the best of our knowledge, MSAD is the first dataset specifically designed for multimodal sentiment analysis in Moroccan dialect, establishing a new benchmark for low-resource Arabic dialect research.

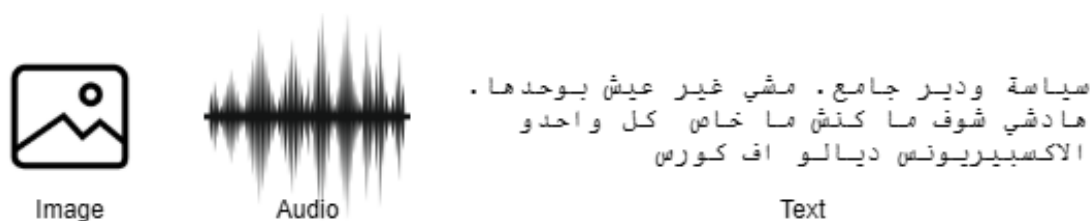


Figure 1. Example data samples from the MSAD-Moroccan dialect dataset

2. METHOD

In this section, the methodology used to construct the MSAD–Moroccan dialect dataset and to establish baseline models for multimodal sentiment analysis is described. Figure 2 illustrates the overall process used to build the dataset and prepare it for multimodal modeling. The dataset is publicly available: <https://data.mendeley.com/datasets/kfjztyzbtb/1>. To develop the proposed dataset and evaluation framework, the workflow is divided into four main stages: data collection from Moroccan podcast videos, multimodal pre-processing and data preparation, manual sentiment annotation, and baseline model construction for multimodal sentiment classification.

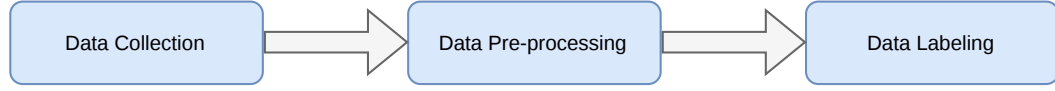


Figure 2. Overview of the dataset creation process

First, the dataset is constructed from authentic Moroccan podcast videos that represent spontaneous conversational speech in Moroccan dialect (Darija). Podcasts are selected as the primary data source because they contain natural discussions accompanied by rich audiovisual cues such as speech intonation, facial expressions, and contextual visual information. In total, approximately 5.13 hours of podcast content are carefully reviewed to identify segments containing clear sentiment expressions. The selected videos are then divided into smaller clips so that each segment corresponds to a coherent sentiment-related utterance. This segmentation process ensures that each instance in the dataset represents a complete multimodal sentiment expression. Each sample is therefore represented as a trimodal instance $M = (T, A, V)$, where T denotes the textual transcript, A represents the audio signal containing prosodic speech features, and V corresponds to the aligned visual representation extracted from the video frames.

A pre-processing pipeline was developed to transform raw podcast videos into structured multimodal data for machine learning. Each video was decomposed into its audio and visual streams, with audio extracted via FFmpeg <https://ffmpeg.org/documentation.html> and resampled to 16 kHz for cross-sample consistency. Audio segments were then temporally aligned with their corresponding video clips. For the visual modality, frames were periodically sampled and passed through OpenCV's Haar Cascade classifier [17], [18] to detect and crop facial regions, which were subsequently resized to 224×224 pixels to ensure uniform input dimensions. Visual features were pre-extracted to optimize computational efficiency, avoiding the storage of raw image data. The textual modality was obtained through automatic transcription of audio segments using OpenAI's Whisper model [19], producing textual sequences that preserve dialectal expressions and code-switching phenomena inherent to Moroccan dialect. The resulting transcripts were temporally aligned with their corresponding audio and visual components, yielding a fully synchronized trimodal dataset.

After pre-processing, the dataset was manually annotated by three native Moroccan speakers to assign sentiment labels (positive, negative, and neutral) to each multimodal sample. Annotations were performed independently, and disagreements were resolved through majority voting. The final label distribution is presented in Table 1, where neutral samples constitute the largest proportion, reflecting the prevalence of ambiguous sentiment in spontaneous conversations.

Table 1. Sentiment label distribution

Sentiment class	Percentage (%)
Positive	28.3
Neutral	42.0
Negative	29.7

Finally, baseline models were developed to evaluate multimodal sentiment classification on the MSAD dataset. Figure 3 presents the overall architecture. Feature representations were extracted independently for each modality using pretrained transformer-based models. MarBERT was employed for textual embeddings tailored to Arabic content [20], HuBERT was used to capture prosodic speech features [21], and ViT extracted high-level visual representations from facial images [22]. These modality-specific encoders enable the extraction of contextual, acoustic, and visual cues relevant to sentiment analysis.

The extracted unimodal feature vectors were used to establish baseline sentiment classification models. In the unimodal setting, each modality was evaluated independently to assess its individual contribution. Bimodal configurations combined modality pairs to examine cross-modal complementarity, while the trimodal setting integrated all three modalities using feature-level fusion prior to classification.

Let $T \in \mathbb{R}^{d_t}$, $A \in \mathbb{R}^{d_a}$, and $V \in \mathbb{R}^{d_v}$ denote the feature representations extracted from the text, audio, and visual modalities, respectively. Two fusion strategies are considered in this work.

First, the concatenation fusion strategy combines the feature vectors by directly concatenating them into a single multimodal representation: $F_{concat} = [T; A; V]$, where $[\cdot]$ denotes vector concatenation. This

approach preserves the full information from each modality and allows the classifier to learn cross-modal relationships in the joint feature space.

Second, a dimensionality-reduction-based fusion method, referred to as *PCA-Sum fusion*, is applied to project each modality into a shared latent space using principal component analysis (PCA). Let P_T , P_A , and P_V represent the PCA projection matrices for the textual, acoustic, and visual modalities. The reduced representations are obtained as: $T' = P_T T$, $A' = P_A A$, $V' = P_V V$, the final fused representation is then computed through element-wise summation: $F_{pca} = T' + A' + V'$. This fusion strategy aims to capture shared latent information across modalities while reducing feature redundancy and mitigating the curse of dimensionality.

For classification, two models are employed: support vector machine (SVM) and multi-layer perceptron (MLP). The SVM classifier is used as a baseline due to its effectiveness in high-dimensional feature spaces and its robustness in small- to medium-sized datasets. In addition, an MLP is used to model non-linear relationships between multimodal features. Given the fused feature vector F , the hidden layer representation is computed as: $h = \sigma(W_1 F + b_1)$, where W_1 and b_1 denote the weight matrix and bias vector of the hidden layer, and $\sigma(\cdot)$ represents the nonlinear activation function (rectified linear unit (ReLU) in our implementation). The output layer then produces sentiment class probabilities using the softmax function: $\hat{y} = \text{softmax}(W_2 h + b_2)$, where $\hat{y}$ represents the predicted probability distribution over the sentiment classes $\{\text{positive}, \text{negative}, \text{neutral}\}$. The final predicted sentiment label is obtained as: $y^* = \arg \max(\hat{y})$. This experimental setup provides a systematic framework for analyzing modality contributions and establishing baseline performance for future multimodal learning approaches on the MSAD–Moroccan dialect dataset.

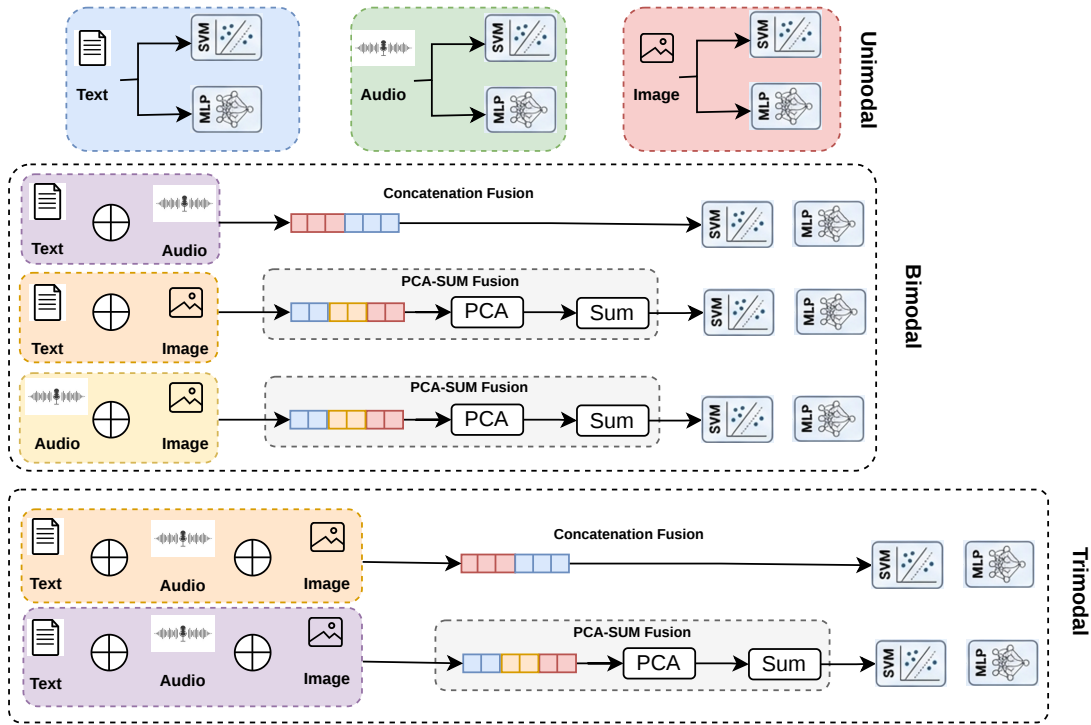


Figure 3. Baseline architecture for unimodal, bimodal, and trimodal sentiment analysis, using modality-specific feature extraction, multimodal fusion, and SVM/MLP classification

3. RESULTS AND DISCUSSION

To assess the effectiveness of MSAD–Moroccan dialect, experiments were conducted across three configurations—unimodal, bimodal, and trimodal—using SVM and MLP classifiers, evaluated on test accuracy and macro F1-score. The experiments examine each modality’s individual contribution and the impact of

multimodal fusion on sentiment prediction performance in Moroccan dialect.

3.1. Unimodal experiments

Unimodal evaluation reveals notable performance differences across modalities. In Table 2, the image modality proved most informative, achieving the highest unimodal accuracy (0.6000) and macro F1-score (0.5805) under the MLP classifier. Audio features yielded moderate results, confirming that prosodic cue—including tone, rhythm, and pitch—offer meaningful sentiment signals in spoken Darija. The textual modality performed the weakest across both classifiers, a result attributable to the linguistic complexity of Moroccan dialect and the prevalence of implicit sentiment expressions that are difficult to capture from text alone.

Table 2. Performance of unimodal sentiment classification using SVM and MLP

Modality	Model	Test accuracy	Macro F1-score
Text	SVM	0.3912	0.3838
Text	MLP	0.4857	0.4497
Audio	SVM	0.4681	0.4556
Audio	MLP	0.5451	0.5267
Image	SVM	0.4418	0.4339
Image	MLP	0.6000	0.5805

Overall, the unimodal results indicate that individual modalities capture only partial information about sentiment expression in real-world conversational content.

3.2. Bimodal experiments

Bimodal experiments evaluated cross-modal combinations using concatenation and PCA-Sum fusion. In Table 3, the Text+Image configuration achieved the best performance (accuracy 0.6198, macro F1-score 0.5975), highlighting strong visual–linguistic complementarity, while Audio+Image showed competitive results and Text+Audio demonstrated limited gains without visual context.

Table 3. Performance of bimodal sentiment classification using two fusion strategies: concatenation and PCA-sum

Modalities	Fusion method	Model	Test accuracy	Macro F1-score
Text+Audio	Concat	SVM	0.4725	0.4656
Text+Audio	Concat	MLP	0.5582	0.5366
Text+Audio	PCA-Sum	SVM	0.4989	0.4953
Text+Audio	PCA-Sum	MLP	0.4857	0.4547
Text+Image	Concat	SVM	0.5209	0.5075
Text+Image	Concat	MLP	0.6198	0.5975
Text+Image	PCA-Sum	SVM	0.5516	0.5400
Text+Image	PCA-Sum	MLP	0.5692	0.5402
Audio+Image	Concat	SVM	0.5121	0.5015
Audio+Image	Concat	MLP	0.5934	0.5581
Audio+Image	PCA-Sum	SVM	0.4923	0.4847
Audio+Image	PCA-Sum	MLP	0.5077	0.4814

3.3. Trimodal experiments

Finally, trimodal experiments were conducted by combining text, audio, and visual modalities simultaneously. The results are summarized in Table 4. The best overall performance was achieved by the MLP classifier using concatenation-based feature fusion, reaching an accuracy of 0.6176 and a macro F1-score of 0.6019. This configuration consistently outperformed all unimodal baselines and most bimodal configurations.

Table 4. Performance of trimodal sentiment classification combining text, audio, and image modalities

Fusion method	Model	Test accuracy	Macro F1-score
Concat	SVM	0.5253	0.5116
Concat	MLP	0.6176	0.6019
PCA-Sum	SVM	0.4857	0.4799
PCA-Sum	MLP	0.5077	0.4870

The improvement observed in the trimodal setting demonstrates that sentiment expression in Moroccan dialect communication is inherently multimodal. By integrating textual content, vocal prosody, and visual cues, the model can capture richer contextual signals that are not available in single-modality systems.

Figures 4 and 5 illustrate the performance trends across different modality combinations and classifiers, considering both accuracy and macro F1-score. The results compare SVM and MLP classifiers under unimodal, bimodal, and trimodal configurations, highlighting the effect of progressively integrating textual, audio, and visual information. Overall, the MLP generally achieves higher performance than SVM across most modality configurations, indicating a stronger ability to capture the nonlinear relationships present in the multimodal features.

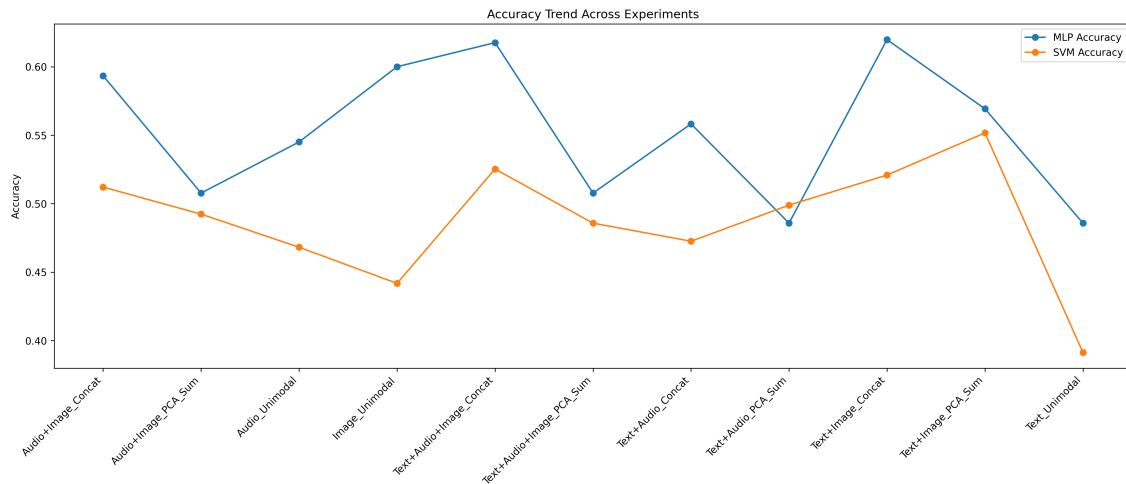


Figure 4. Accuracy comparison across unimodal, bimodal, and trimodal experiments for SVM and MLP classifiers on the MSAD–Moroccan dialect dataset

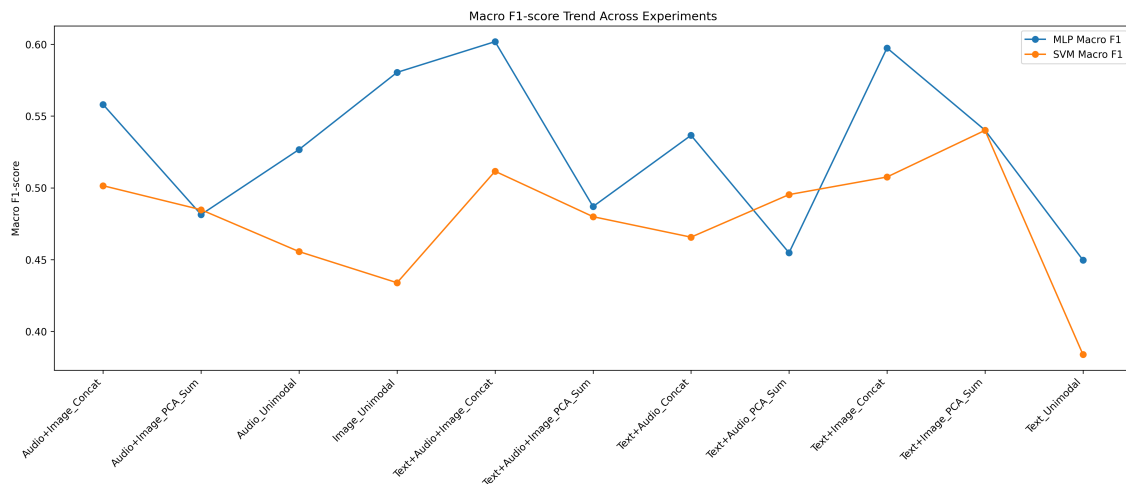


Figure 5. Macro F1-score comparison across different modality configurations and fusion strategies

3.4. Discussion

The experimental results provide several important insights regarding multimodal sentiment analysis for Moroccan dialect. First, multimodal fusion consistently outperforms unimodal approaches in sentiment classification. This finding aligns with prior studies emphasizing that human sentiment expression relies on complementary linguistic, acoustic, and visual cues [12], [13]. The improvement observed in bimodal and trimodal configurations confirms the complementary nature of these modalities. Second, visual information

plays a critical role in sentiment recognition, as unimodal and bimodal results show significant performance gains from the image modality. Third, existing Moroccan dialect datasets, such as MAC [23] and MSTD [24], primarily focus on textual data and overlook the multimodal nature of real-world communication. Although MDER-MA [25] introduces multimodal data, it targets emotion recognition rather than sentiment classification, leaving multimodal sentiment analysis in Moroccan dialect underexplored.

The MSAD–Moroccan dialect dataset fills an important gap by providing the first multimodal resource dedicated to sentiment classification in Moroccan dialect. Unlike existing datasets that focus mainly on textual social media data, MSAD integrates synchronized textual, acoustic, and visual modalities extracted from authentic podcast videos. This multimodal design offers a more realistic representation of sentiment expression in natural conversations. By incorporating vocal prosody and facial cues alongside text, the dataset enables the development of models that better capture the nuanced and context-dependent nature of sentiment in spoken Moroccan dialect.

The experimental results confirm the importance of multimodal fusion for sentiment analysis in Moroccan dialect media. Unimodal models capture only partial sentiment cues, whereas bimodal and trimodal configurations consistently achieve better performance. In particular, the combination of textual and visual modalities demonstrates strong complementarity, indicating that facial expressions and contextual imagery enhance linguistic interpretation. The best results are obtained with trimodal concatenation-based fusion using an MLP classifier, achieving an accuracy of 0.6176 and a macro F1-score of 0.6019. These findings emphasize the advantage of integrating heterogeneous modalities. Furthermore, the dataset has broader implications for embedded systems, reconfigurable computing, and internet of thing (IoT) applications, where efficient multimodal processing is essential for deployment in resource-constrained environments.

4. CONCLUSION

This work introduces MSAD-Moroccan dialect, a multimodal dataset designed for sentiment analysis in Moroccan dialect speech. It provides aligned textual, acoustic, and visual data from authentic podcast conversations annotated by native speakers, enabling evaluation of unimodal and multimodal models. Results show that multimodal fusion significantly improves performance over unimodal approaches. Despite its contributions, limitations include modest dataset size, class imbalance, and reliance on baseline models. Future work will expand the dataset, enrich annotations, and explore advanced multimodal architectures and real-time applications.

FUNDING INFORMATION

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Ayoub Ben Cheikhi	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓		✓	✓
El Habib Nfaoui	✓	✓		✓	✓			✓		✓	✓	✓		✓

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal Analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project Administration

Fu : Funding Acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest

DATA AVAILABILITY

The dataset supporting the findings of this study is publicly available and can be accessed online. All data used and analyzed during the current study are available at the following repository: <https://data.mendeley.com/datasets/kfjztyzt>

REFERENCES

- [1] P. Kaifosh and T. R. Reardon, "A generic non-invasive neuromotor interface for human-computer interaction," *Nature*, vol. 645, no. 8081, pp. 702–711, 2025, doi: 10.1038/s41586-025-09255-w.
- [2] R. Pillalamarri and U. Shanmugam, "A review on EEG-based multimodal learning for emotion recognition," *Artificial Intelligence Review*, vol. 58, no. 5, p. 131, 2025, doi: 10.1007/s10462-025-11126-9.
- [3] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 478–493, 2020, doi: 10.1109/JSTSP.2020.2987728.
- [4] S. Jabeen, X. Li, M. S. Amin, O. Bourahla, S. Li, and A. Jabbar, "A review on methods and applications in multimodal deep learning," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 19, no. 2s, 2023, doi: 10.1145/3545572.
- [5] M. A. Almulla, "A multimodal emotion recognition system using deep convolution neural networks," *Journal of Engineering Research*, vol. 13, no. 2, pp. 721–729, 2025, doi: 10.1016/j.jer.2024.03.021.
- [6] R. Gummula, V. Arumugam, and A. Aranganathan, "Multimodal recognition with deep learning: Audio, image, and text," *International Journal of Reconfigurable and Embedded Systems (IJRES)*, vol. 14, no. 1, pp. 254–264, 2025, doi: 10.11591/ijres.v14.i1.pp254-264.
- [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, doi: 10.1016/j.inffus.2022.09.025.
- [8] A. Ligthart, C. Catal, and B. Tekinerdogan, "Systematic reviews in sentiment analysis: A tertiary study," *Artificial Intelligence Review*, vol. 54, pp. 4997–5053, 2021, doi: 10.1007/s10462-021-09973-3.
- [9] G. Xu, Y. Meng, X. Qiu, Z. Yu, and X. Wu, "Sentiment analysis of comment texts based on BiLSTM," *IEEE Access*, vol. 7, pp. 51522–51532, 2019, doi: 10.1109/ACCESS.2019.2909919.
- [10] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731–5780, 2022, doi: 10.1007/s10462-022-10144-1.
- [11] K. Zhao, M. Zheng, Q. Li, and J. Liu, "Multimodal sentiment analysis—A comprehensive survey from a fusion methods perspective," *IEEE Access*, vol. 13, pp. 64556–64583, 2025, doi: 10.1109/ACCESS.2025.3554665.
- [12] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, and X. Kong, "Multimodal sentiment analysis based on fusion methods: A survey," *Information Fusion*, vol. 95, pp. 306–325, 2023, doi: 10.1016/j.inffus.2023.02.028.
- [13] S. Lai, X. Hu, H. Xu, Z. Ren, and Z. Liu, "Multimodal sentiment analysis: A survey," *Displays*, vol. 80, 2023, doi: 10.1016/j.displa.2023.102563.
- [14] A. B. Cheikhi and E. H. Nfaoui, "GCMA-Net: A gated cross-modal attention network for Arabic multimodal sentiment analysis," *IEEE Access*, vol. 14, pp. 30733–30746, 2026, doi: 10.1109/ACCESS.2026.3667314.
- [15] 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)*, 2023, pp. 126–133, doi: 10.48550/arXiv.2306.06322.
- [16] 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*, 2017, pp. 2378–2386, doi: 10.1007/978-981-10-6571-2_290.
- [17] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.
- [18] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, Kauai, HI, USA, 2001, p. I, doi: 10.1109/CVPR.2001.990517.
- [19] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*, 2023, pp. 28492–28518.
- [20] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep bidirectional transformers for Arabic," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, 7088–7105, doi: 10.18653/v1/2021.acl-long.551.
- [21] W.-N. Hsu, B. Bolte, Y.-H. Tsai, K. Lakhota, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021, doi: 10.1109/TASLP.2021.3122291.
- [22] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv Preprint*, 2021, doi: 10.48550/arXiv.2010.11929.
- [23] M. Garouani and J. Kharroubi, "MAC: An open and free Moroccan Arabic corpus for sentiment analysis," in *The Proceedings of the International Conference on Smart City Applications*, 2021, pp. 849–858, doi: 10.1007/978-3-030-94191-8_68.
- [24] S. Mihi, B. Ait, I. El, S. Arezki, and N. Laachfoubi, "MSTD: Moroccan sentiment Twitter dataset," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 10, pp. 363–372, 2020, doi: 10.14569/IJACSA.2020.0111045.
- [25] S. Ouali and S. El Garouani, "MDER-MA: a multimodal dataet for emotion recognition in low-resource Moroccan Arabic language," *Data in Brief*, vol. 62, 2025, doi: 10.1016/j.dib.2025.112005.

BIOGRAPHIES OF AUTHORS

Ayoub Ben Cheikhi    is a third-year Ph.D. student in Computer Science, specializing in artificial intelligence and multimodal applications using pretrained models and LLMs. With a strong academic background—including a Master's degree in Decision Science and Intelligent Vision. He is currently conducting research on multimodal applications using pretrained models and LLMs, large language models, deep learning, and image processing at the L3IA Laboratory, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University. He has contributed to several international publications on topics such as multimodal sentiment analysis and sarcasm detection. He can be contacted at email: ayoub.bencheikhi@usmba.ac.ma.



El Habib Nfaoui    is currently a Professor of Computer Science at the University of Sidi Mohamed Ben Abdellah, Fez, Morocco. He received his Ph.D. in Computer Science from the University of Sidi Mohamed Ben Abdellah, Morocco, and the University of Lyon, France, under a Co-tutelle agreement (doctorate in joint supervision), in 2008, and then his HU Diploma (Accreditation to supervise research) in Computer Science, in 2013, from the University of Sidi Mohamed Ben Abdellah. His current research interests include natural language processing, large language models, web mining and text mining, information retrieval, machine learning and deep learning, semantic web, web services, social networks, and multi-agent systems. He has supervised 16 Ph.D. theses related to his research interests. He has published in international reputed journals, books, and conferences, and has edited conference proceedings and journal special issues. He has served as a reviewer for scientific journals and as a program committee of several conferences. He has served as an expert at the National Center for Scientific and Technical Research (CNRST) and Mohammed VI Polytechnic University (UM6P), Morocco. He is co-founder and Chair of the IEEE Morocco Section Computational Intelligence Society Chapter. He is a co-founder and an executive member of the International Neural Network Society Morocco regional chapter. He co-founded the International Conference on Intelligent Computing in Data Sciences (ICSD2017) and the International Conference on Intelligent Systems and Computer Vision (ISCV2015). He can be contacted at email: elhabib.nfaoui@usmba.ac.ma.