L3IA-Subtask 1 at AraSentEval Shared Task: Multi-Dialect Arabic Sentiment Classification via a Transformer-Based Approach

Mohamed M'haouach, Kaouthar Elyoussoufi, Hamza Alami5, Abdessamad Benlahbib
L3IA Laboratory, Faculty of Sciences Dhar El Mahraz,
Sidi Mohamed Ben Abdellah University, Fez, 30003, Morocco
{mohamed.mhaouach, kaouthar.elyoussoufi, hamza.alami5,
abdessamad.benlahbib}@usmba.ac.ma

Abstract

This paper presents our system and findings for AraSentEval 2026 Subtask 1 on Arabic Dialect Sentiment Analysis. We propose an automated sentiment classification system grounded in advanced Natural Language Processing (NLP) techniques. The proposed approach leverages pre-trained Transformer-based architectures to categorize textual inputs into three sentiment polarities: positive, negative, and neutral. Initially, a text normalization procedure is applied to unify the orthographic and graphical variations characteristic of the Arabic language. This process is further complemented by repetition reduction techniques, which aim to mitigate textual noise and enhance the overall consistency of the data. Subsequently, the data are adapted to the requirements of the pre-trained models to ensure coherent tokenization. The processed texts are then encoded into numerical representations that serve as inputs during training and evaluation. Finally, we conduct a comprehensive benchmarking study of five Transformer-based architectures to assess their effectiveness. The best-performing experimental setup yielded remarkable results on the AraSentEval 2026 benchmark, achieving a micro-F1 score of 75.96% on the official test set.

Keywords: Sentiment analysis, machine learning, Arabic dialects

1. Introduction

Sentiment analysis in Arabic dialects constitutes a major challenge in Natural Language Processing (NLP). This challenge arises in a context shaped by the rapid expansion of web applications and social media platforms, which continuously generate large volumes of reviews and user-generated comments across diverse domains(Matrane et al., 2023). Although these data represent a valuable resource for large-scale opinion and emotion analysis, their effective exploitation in the Arabic context remains difficult due to the coexistence of Modern Standard Arabic (MSA) and multiple country-specific dialects. Unlike MSA, which benefits from a relatively standardized linguistic system, Arabic dialects lack orthographic normalization and exhibit substantial lexical, syntactic, and morphological variability. Consequently, this linguistic heterogeneity limits the ability of computational models to learn truly generalizable representations and may undermine the robustness of sentiment predictions.

In recent years, increasing attention has been devoted to incorporating Arabic dialects into sentiment analysis within the field of Natural Language Processing. Nevertheless, the majority of existing studies remain primarily focused on Modern Standard Arabic (MSA)(Al-Harbi, 2019). As a consequence, models mainly trained on MSA often struggle to generalize to the wide range of dialectal varieties, which may undermine the reliability and robustness of predictions in real-world settings. In this context, the AraSentEval 2026 shared task (Ezzini et al., 2026; Alharbi et al., 2025a,c,b) introduces a common experimental framework for sentiment analysis across multiple Arabic dialects. Subtask 1 addresses this challenge as a multi-class classification problem (positive, negative, neutral), requiring systems to effectively handle dialectal linguistic variability despite the scarcity of high-quality annotated resources and the inherent linguistic heterogeneity that characterizes these varieties.

In this paper, we present our system submitted to the AraSentEval 2026 shared task, which aims to advance sentiment classification techniques across Arabic dialects within the hospitality domain. To encode the input texts, we investigate the use of several pre-trained language models, namely BERT(Devlin et al., 2019) , AraBERT(Antoun et al., 2020), MARBERT(Abdul-Mageed et al., 2021), CAMeLBERT(Inoue et al., 2021), and DarijaBERT(Gaanoun et al., 2025). Each model is fine-tuned under identical experimental settings to ensure a fair and systematic comparison. This comparative study enables us to identify the most effective architecture for sentiment classification in dialect-rich Arabic data. Furthermore, it provides empirical insights into the robustness and generalization capacity of Transformer-based models when deployed in multi-dialectal environments.

This paper is organized as follows. Section 2 presents the proposed method and describes the overall approach adopted in this work. Section 3 reports the experimental setup and discusses the obtained results. Finally, Section 4 concludes the paper by summarizing the main findings.

Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks, pages 292–295
May 11, 2026.

2. Method

This section briefly outlines the methodology adopted for the development of an Arabic Dialect Sentiment Classification system. The proposed framework follows a structured pipeline, spanning from data preparation to the final classification stage, with the overall training procedure illustrated in Figure 1.

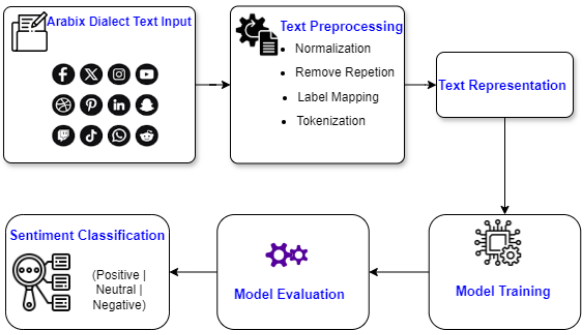

Figure 1: Flowchart of the Sentiment Classification System

2.1. Arabic Dialect Text Preprocessing

For the processing of Arabic dialectal texts, a lightweight preprocessing pipeline was employed to ensure compatibility with the underlying models while preserving the intrinsic linguistic information. An orthographic normalization step was introduced to standardize graphical variants and reduce formal variability that could otherwise amplify lexical sparsity. In addition, a label mapping procedure was performed, whereby the textual sentiment labels ("positive", "negative", "neutral") were transformed into numerical identifiers (0, 1, 2), thus enabling their effective integration within supervised learning algorithms. Finally, each review was tokenized using the model-specific pre-trained tokenizer, thereby ensuring strict alignment between the input representation and the specifications of the corresponding pre-training phase.

2.2. Arabic Dialect Text Representation

To represent Arabic dialect texts, we employ contextualized embeddings generated by pre-trained Transformer architectures. Given an input sequence $X = (x_1, \ldots, x_n)$, tokens are first mapped into dense embeddings combining token, positional, and segment information, and then processed through stacked self-attention layers to capture contextual and dialect-specific patterns. For classification, the final hidden state of the special token `[CLS]`, denoted as h_{CLS}, is used as a global representation of the text and fed into the downstream sentiment classifier.

2.3. Baseline Classifiers

In this study, we relied on five pre-trained Transformer-based architectures obtained from the HuggingFace Transformers library as our foundational models, namely: BERT, AraBERT, MARBERT, CAMeLBERT, and DarijaBERT.

- **BERT**: A pre-trained Transformer-based language model developed by Google AI, designed to learn bidirectional contextual representations for various downstream NLP tasks
- **AraBERT**: A pre-trained Arabic Transformer model designed to achieve strong performance across multiple Arabic NLP benchmarks.
- **MARBERT**: is an Arabic Transformer-based language model pre-trained on large and diverse corpora to enable transfer learning for both Modern Standard Arabic and dialectal Arabic.
- **CAMeLBERT**: is a suite of BERT-based models pre-trained on Arabic corpora, available in various sizes and configurations to support diverse NLP applications.
- **DarijaBERT**: is a suite of language models built upon the BERT architecture, specifically developed to address natural language processing tasks in the Moroccan Arabic dialect Darija.

3. Experimental Results

In this study, we conduct our experimental evaluation on AraSentEval 2026 – Subtask 1: Arabic Dialect Sentiment Analysis. All experiments were performed in the Google Colab environment[1] to ensure computational efficiency and reproducibility. Model implementation and fine-tuning were carried out using the Hugging Face Transformers library[2] (Wolf et al., 2020) , while PyTorch[3] was employed to handle the training procedure and performance evaluation. This experimental framework allows for a rigorous and systematic assessment of the proposed models under standardized and reproducible conditions.

3.1. Datasets

The experimental evaluation was conducted on the Multi-Dialect-Sent (MDS-3) dataset, a benchmark designed for sentiment analysis in Arabic dialects.

[1] https://colab.research.google.com/
[2] https://huggingface.co/docs/transformers/index
[3] https://pytorch.org/

The corpus consists of 3,000 sentences annotated with three sentiment labels: positive, negative, and neutral. The data were collected from hotel reviews and subsequently translated and manually annotated by native speakers of the corresponding dialects. For the experiments, the dataset was divided into training (85%) and validation (15%) sets using label stratification to preserve the original class distribution. Figures 2 and 3 illustrate the distribution of sentiment categories and the distribution of Arabic dialects within the corpus.

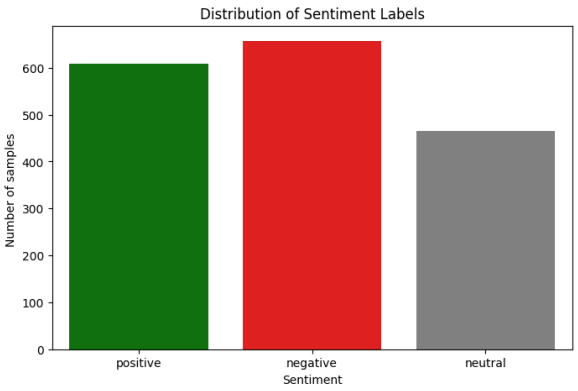

Figure 2: Distribution of Sentiment Labels

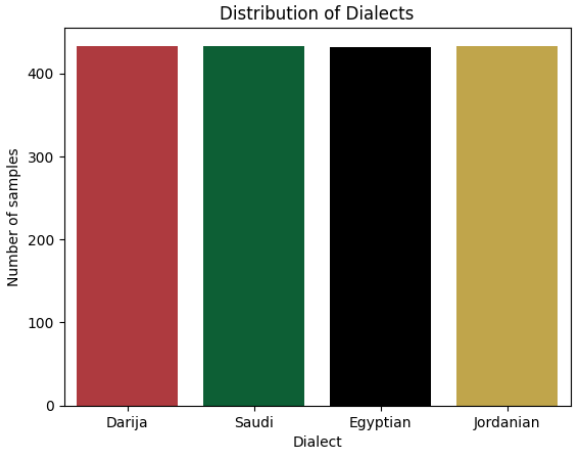

Figure 3: Distribution of Dialects

3.2. Experimental Settings

During the fine-tuning phase, all Transformer-based models were trained under a unified experimental configuration. The optimization process relied on the AdamW optimizer with a learning rate of 2×10^{-5} and a weight decay of $\lambda = 0.01$. The training procedure was conducted for 5 epochs using a batch size of 16. Model parameters were optimized using the Cross-Entropy loss function to handle the multi-class sentiment classification task.

3.3. Performance Evaluation

In this section, we present a comparative evaluation of the different models trained for Subtask 1. Their performance is assessed using standard classification metrics to measure their effectiveness in identifying sentiment polarity across multiple Arabic dialects.

In this experiment, we evaluate and compare several Transformer-based models for the task of Arabic dialect sentiment classification, as reported in Table 1 and illustrated in Figure 4. The results clearly indicate that MARBERT achieves the best overall performance, obtaining the highest Accuracy (95.10%) and F1-score (94.92%), which highlights its strong ability to model dialectal variations. AraBERT follows with competitive results (93.08% Accuracy and 92.54% F1-score), demonstrating balanced predictive performance. CAMeLBERT and DarijaBERT achieve slightly lower yet comparable performances, with F1-scores of 90.16% and 90.45%, respectively. In contrast, BERT records the lowest performance (88.50% F1-score), likely due to its general-purpose pre-training not being specifically optimized for Arabic dialectal data.

Table 1: Performance models

Model	Accuracy	Precision	Recall	F1-score
BERT	88.76%	88.37%	88.69%	88.50%
AraBERT	93.08%	93.52%	92.12%	92.54%
MARBERT	95.10%	94.95%	94.92%	94.92%
CAMeLBERT	90.49%	90.08%	90.25%	90.16%
DarijaBERT	90.78%	90.97%	90.32%	90.45%

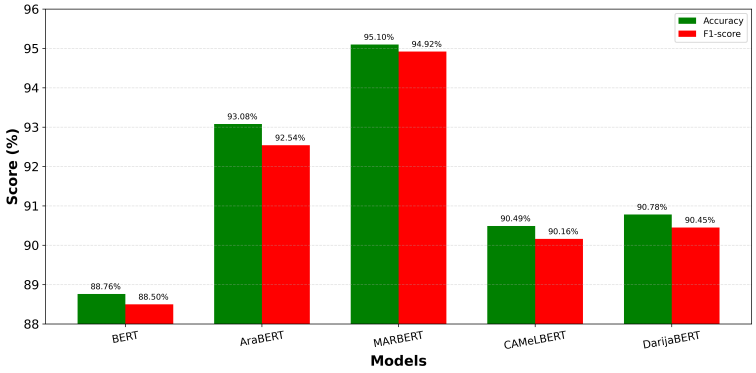

Figure 4: Comparative Performance of the Models

4. Conclusion

This work presented our system for AraSentEval 2026 Subtask 1, which addresses sentiment classification across diverse Arabic dialects. The proposed approach relies on a Transformer-based framework combined with a lightweight preprocessing pipeline, including orthographic normal-

ization and label mapping. Several pre-trained language models were evaluated in order to determine their effectiveness for multi-dialect sentiment analysis. Experimental results on the Multi-Dialect-Sent (MDS-3) dataset show that Transformer-based models can effectively capture sentiment polarity in dialect-rich Arabic texts. Among the evaluated architectures, MARBERT achieved the best performance, reaching an accuracy of 95.10% and a macro F1-score of 94.92%, highlighting the advantage of dialect-aware pre-trained models for handling linguistic variability in Arabic dialects.

5. References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. 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 (ACL-IJCNLP 2021), Volume 1: Long Papers*, pages 7088–7105. Association for Computational Linguistics.

Omar Al-Harbi. 2019. Classifying sentiment of dialectal arabic reviews: A semi-supervised approach. *International Arab Journal of Information Technology*.

Maram Alharbi, Saad Ezzini, Hansi Hettiarachchi, Tharindu Ranasinghe, and Ruslan Mitkov. 2025a. Evaluating large language models on arabic dialect sentiment analysis. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Maram I Alharbi, Salmane Chafik, Saad Ezzini, Ruslan Mitkov, Tharindu Ranasinghe, and Hansi Hettiarachchi. 2025b. Ahasis: Shared task on sentiment analysis for arabic dialects. In *Proceedings of the Shared Task on Sentiment Analysis for Arabic Dialects*, pages 1–6.

Maram I Alharbi, Saad Ezzini, Hansi Hettiarachchi, Tharindu Ranasinghe, and Ruslan Mitkov. 2025c. Ador: Dataset for arabic dialects in hotel reviews: A human benchmark for sentiment analysis. In *Proceedings of the First Workshop on Advancing NLP for Low-Resource Languages*, pages 187–191.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT 2020), with a Shared Task on Offensive Language Detection*, pages 9–15. European Language Resources Association (ELRA).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, pages 4171–4186, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Saad Ezzini, Shadi Abudalfa, Maram Alharbi, Salmane Chafik, Samaher Alghamdi, Reem Alotaibi, Hamzah Luqman, Mo El-Haj, and Paul Rayson. 2026. AraSentEval: A shared task on sentiment analysis and swapping in arabic. In *Proceedings of the 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

Kamel Gaanoun, Abdou Mohamed Naira, Anass Allak, and Imade Benelallam. 2025. Darijabert: A step forward in nlp for the written moroccan dialect. *International Journal of Data Science and Analytics*, 20(2):917–929.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop (WANLP 2021)*, pages 92–104, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Youssef Matrane et al. 2023. A systematic literature review of arabic dialect sentiment analysis. *Journal of King Saud University – Computer and Information Sciences*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Jamie Brew, Canwen Huang, HuggingFace Inc., et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.