LinguArabic at AraSentEval Shared Task: MARBERT for Multi-Dialect Arabic Sentiment Analysis

Norah Saud Alshahrani,Elham Abdullah Al-Qarni ,Shatha Hussan Alshomrani
Najran University, Saudi Arabia
Al-Baha Private College of Sciences, Saudi Arabia
Independent Researcher, Saudi Arabia
nsnalshahrani@nu.edu.sa, elhamalqarni@bpcs.edu.sa, shatha2030ksa@gmail.com

Abstract

Sentiment analysis for Arabic dialects remains challenging due to substantial linguistic variation across dialects and the expansion of informal language in user-generated content. The AraSentEval 2026 shared task introduces a multi-dialect benchmark designed to evaluate sentiment classification systems on real-world Arabic data. In this paper, we present LinguArabic's submission to the sentiment classification track of AraSentEval 2026. Our approach is based on fine-tuning MARBERT, a transformer model pre-trained on large-scale Arabic social media data that captures diverse dialectal patterns. To improve model robustness, we incorporate a multi-stage preprocessing pipeline that includes text normalization, dialect-aware lexical mapping, and confidence-based prediction adjustment. We specifically investigate the impact of advanced normalization rules in reducing lexical sparsity across various regional dialects. Experimental results show that the proposed system achieves a Macro F1-score of 0.8333 on the official evaluation set. Our findings highlight the importance of dialect-aware pretraining and preprocessing strategies for improving sentiment classification performance across diverse Arabic dialects, providing a scalable framework for real-world Arabic NLP applications.

Keywords: Sentiment Analysis, Arabic Dialect Sentiment Analysis, MARBERT.

1. Introduction

Sentiment Analysis (SA) is a core task in Natural Language Processing (NLP) that aims to automatically identify opinions, attitudes, and emotions expressed in text. In Arabic, sentiment analysis has received increasing attention, with efforts focusing on the development of sentiment lexicons, benchmark datasets, and shared tasks (El-Beltagy et al., 2017). However, most early work primarily targets Modern Standard Arabic (MSA), which does not accurately reflect the linguistic characteristics of user-generated online content.

Arabic dialects pose significant challenges for sentiment analysis due to their high variability in spelling, vocabulary, morphology, and syntax. Unlike MSA, dialects are rarely standardized and often include informal expressions, phonetic writing, and code-switching. Since most social media posts and online reviews are written in dialect, models trained on formal Arabic frequently struggle to generalize effectively.

To address this limitation, the AraSentEval 2026 shared task introduces a multi-dialect benchmark for sentiment classification and sentiment swapping in real-world Arabic data (Ezzini et al., 2026). The dataset builds upon recent resources such as ADOR (Alharbi et al., 2025c) and follows previous shared-task efforts including AHASIS (Alharbi et al., 2025b). By providing a multi-dialect dataset, AraSentEval encourages the development of models that better handle dialectal diversity in user-generated content.

In this paper, we describe LinguArabic's submission to the sentiment classification track of AraSentEval 2026. Our system is based on MARBERT, a transformer model pre-trained on large-scale Arabic social media data containing diverse dialects (Abdul-Mageed et al., 2021). We fine-tune MARBERT on the provided dataset using stratified data splitting, dialect-aware preprocessing, and macro-F1–based model selection. Our results show that dialect-aware pretraining plays an important role in improving sentiment classification across Arabic dialects.

2. Background

2.1. Task Description

The AraSentEval shared task focuses on sentiment analysis for Arabic dialects, aiming to evaluate models on realistic user-generated dialectal text. In Subtask 1 (Arabic Dialect Sentiment Analysis), the objective is to determine the sentiment polarity expressed in a dialectal Arabic sentence. The system receives an input sentence x written in a specific Arabic dialect and predicts a sentiment label $y \in \{\textit{positive}, \textit{negative}, \textit{neutral}\}$.

2.2. Dataset Description

Our experiments were conducted using the Multi-Dialect-Sent (MDS-3) dataset provided for the

Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks, pages 302–305
May 11, 2026.

AraSentEval shared task. The dataset contains 2,043 sentences annotated with three sentiment labels: *positive*, *negative*, and *neutral*. It covers four major Arabic dialects—Moroccan, Egyptian, Jordanian, and Saudi—allowing evaluation across multiple regional varieties of Arabic.

The sentences were collected from hotel reviews, representing user-generated content commonly used in sentiment analysis research (Alharbi et al., 2025c). The reviews were translated and manually annotated by native speakers of each dialect to ensure accurate sentiment labeling.

The test set, released during the evaluation phase, contains 312 dialectal reviews from the same dialect groups. Participants were required to predict the sentiment polarity of each sentence using their trained models and submit the predictions to the shared task platform for automatic evaluation. The results were displayed on a public leaderboard showing the performance of all participating systems.

2.3. Related Work

Arabic sentiment analysis has been extensively studied in recent years. Early research primarily focused on Modern Standard Arabic (MSA) and relied on sentiment lexicons or traditional machine learning methods (El-Beltagy et al., 2017). Shared tasks such as SemEval Sentiment Analysis in Twitter (Rosenthal et al., 2017) further advanced benchmark development and standardized evaluation methodologies.

Despite this progress, dialectal Arabic sentiment analysis remains challenging due to orthographic variation, lexical diversity, and the limited availability of large annotated datasets. Recent resources such as ADOR (Alharbi et al., 2025c) and dialect-focused shared tasks like AHASIS (Alharbi et al., 2025b) have emphasized the importance of evaluating models on dialectal data. Recent studies have also investigated the effectiveness of large language models for Arabic dialect sentiment analysis, highlighting both the potential and limitations of such models when dealing with dialectal variation (Alharbi et al., 2025a).

In contrast to earlier approaches centred on MSA, our work leverages MARBERT, a transformer model pre-trained on large-scale Arabic social media data containing substantial dialectal variation (Abdul-Mageed et al., 2021). By fine-tuning MARBERT on the MDS-3 dataset, we aim to better capture dialect-specific linguistic patterns and improve sentiment classification across multiple Arabic dialects.

While transformer-based models like AraBERT have set benchmarks for Modern Standard Arabic (MSA) (Antoun et al., 2020), MARBERT's specialized pre-training on social media data allows it to better capture the linguistic nuances of dialectal Arabic, which is essential for this task.

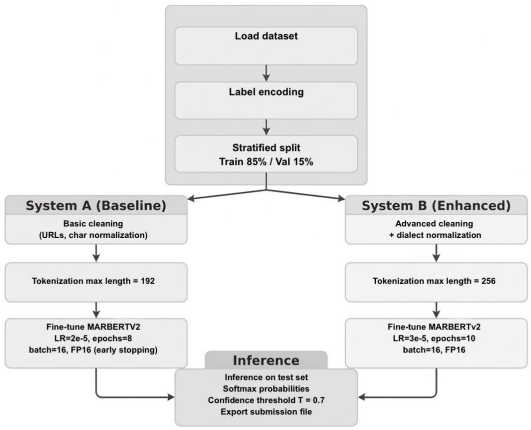

Figure 1: Overview of the proposed MARBERT-based system for Arabic dialect sentiment analysis.

3. System Overview

Our system is based on MARBERTv2, a transformer-based language model designed for Arabic dialects (Abdul-Mageed et al., 2021). MARBERT follows the *BERT*$_{base}$ architecture with 12 transformer layers, 12 self-attention heads, and a hidden size of 768. It was pre-trained using the masked language modeling objective on more than one billion Arabic tweets containing both Modern Standard Arabic (MSA) and multiple dialects.

For sentiment classification, we fine-tune MARBERT on the MDS-3 dataset by adding a fully connected classification layer on top of the encoder. The contextual representation of the `[CLS]` token is used as the sentence-level representation. Given an input sentence $x = (w_1, \ldots, w_n)$, the encoder produces a contextual embedding $h = \text{MARBERT}(x)_{[\text{CLS}]}$, and the final prediction is computed as $y = \text{softmax}(Wh + b)$, where W and b denote the classifier parameters.

To improve performance on dialectal Arabic, we apply a preprocessing pipeline that removes URLs, normalizes Arabic characters, reduces repeated characters, removes punctuation, and performs rule-based dialect normalization. The cleaned text is then tokenized using the MARBERT tokenizer with a maximum sequence length of 192 tokens in the baseline system and 256 tokens in the enhanced configuration. The overall experimental pipeline and system configurations are illustrated in Figure 1.

Dialect	Normalization Examples
Moroccan	جدا → بزاف ;جيد → مزيان، زين ;هذا → هاد، هاذ، هدا
Egyptian	ايه، ايوه ;هكذا → كده، كدا ;جدا → اووي، كتير، كتير → نعم
Gulf	دائماً → دايم ;الآن → الحين، دحين ;جدا → مره، مرة
Multiple	لي، ;لايوجد → ماكاينش، مفيش ;موجود → كاين، كين الذي → اللي

Table 1: Examples of dialect normalization rules used in the preprocessing pipeline.

4. Experimental Setup

The system was implemented using PyTorch and the HuggingFace Transformers library. MARBERTv2 was loaded from the HuggingFace repository (`UBC-NLP/MARBERTv2`). Experiments were conducted on Google Colab using GPU acceleration and mixed precision training (FP16).

4.1. Preprocessing and Tokenization

Arabic dialectal text often contains noisy patterns such as URLs, repeated characters, and spelling variations. We therefore apply a preprocessing pipeline that removes URLs, normalizes Arabic characters (including the letter ى → ي), removes punctuation and special characters, and reduces repeated characters. These steps reduce lexical variation while preserving dialectal expressions.

To further improve model robustness, the enhanced system applies rule-based normalization of dialectal variants. Examples of these mappings are shown in Table 1.

4.2. Training Configuration

The model was fine-tuned using the AdamW optimizer with a learning rate of 3×10^{-5} and a batch size of 16. Mixed-precision training (FP16) was used to improve training efficiency. To prevent overfitting, early stopping with a patience of two epochs was applied, and the best-performing model was selected based on validation loss.

4.3. System Configurations

Two system configurations were evaluated.

System A (Baseline MARBERT). This configuration fine-tunes MARBERT using basic preprocessing and a maximum sequence length of 192 tokens.

System B (Enhanced MARBERT). This configuration incorporates advanced preprocessing techniques including dialect normalization rules, a longer sequence length (256 tokens), and confidence-based prediction filtering.

4.4. Confidence-Based Prediction Adjustment

To reduce uncertain predictions, a confidence-based filtering strategy was applied during inference. After computing the softmax probabilities, the maximum prediction confidence was calculated. If the confidence score fell below a predefined threshold τ, the prediction was reassigned to the *neutral* class (Guo et al., 2017). Formally:

$$y_{\text{final}} = \begin{cases} \text{neutral} & \text{if } \max(p) < \tau \\ \arg\max(p) & \text{otherwise} \end{cases}$$

To determine the value of the confidence threshold τ, we conducted preliminary experiments on the validation set of the MDS-3 dataset. Several threshold values in the range $[0.5, 0.9]$ were evaluated, and $\tau = 0.7$ produced the most stable performance in terms of Macro F1-score.

4.5. Evaluation Metrics

Model performance was evaluated using standard multi-class classification metrics. The primary evaluation metric of the shared task is the **Macro F1-score**, which computes the F1-score independently for each sentiment class and then averages the results. Additional metrics reported include accuracy, precision, recall, and weighted F1-score.

5. Results

The proposed system was evaluated on the official AraSentEval test set using the shared task evaluation platform. Model performance was measured using Macro F1-score, the primary evaluation metric for the task, along with accuracy, precision, and recall.

Our MARBERT-based system achieved a Macro F1-score of 0.8333, while accuracy, precision, and recall also reached 0.8333, indicating balanced performance across the three sentiment classes.

To analyze the impact of preprocessing strategies, we compared two system configurations. System A applies MARBERT with basic preprocessing, while System B introduces dialect normalization, advanced preprocessing, and confidence-based prediction filtering. The comparison results are presented in Table 2. The enhanced configuration achieves higher performance, demonstrating improved robustness when handling dialectal lexical variation.

Error analysis shows that most misclassifications occur between the neutral and positive classes. For example, the sentence الفندق عادي والخدمة مقبولة (“The hotel is average and the service is acceptable”) was labeled as neutral but pre-

System	Description	Macro F1
System A	MARBERT + basic preprocessing	0.80
System B	MARBERT + dialect normalization + confidence filtering	0.83

Table 2: Comparison of system configurations for dialect sentiment classification.

dicted as positive due to the mildly positive word مقبولة.

Another challenge involves implicit or sarcastic sentiment expressions. In the sentence الخدمة ممتازة إذا كنت تحب الانتظار ساعتين ("The service is excellent if you enjoy waiting two hours"), the phrase الخدمة ممتازة contains strong positive cues, although the overall meaning expresses dissatisfaction.

Dialectal vocabulary variation also contributes to errors. For instance, المكان مزيان ولكن السعر بزاف ("The place is good but the price is too high") includes dialectal terms such as مزيان and بزاف that require contextual interpretation.

Overall, the results show that transformer-based models such as MARBERT are effective for Arabic dialect sentiment analysis. However, challenges remain in handling sarcasm, ambiguous neutral expressions, and dialectal lexical diversity.

6. Conclusion

In this paper, we presented our system for the AraSentEval shared task on Arabic dialect sentiment analysis. Our approach fine-tunes the **MARBERTv2** transformer model, which is pre-trained on large-scale Arabic social media data containing diverse dialectal varieties. To improve robustness when processing dialectal text, we applied a preprocessing pipeline including text normalization, dialect-aware lexical mapping, and confidence-based prediction adjustment.

Experimental results show that our system achieves a **Macro F1-score of 0.8333** on the official evaluation set. The comparison of system configurations demonstrates that dialect normalization and advanced preprocessing strategies improve classification performance by reducing lexical sparsity and enabling richer contextual representations.

Error analysis reveals several remaining challenges, including sarcasm detection, ambiguous neutral expressions, and dialect-specific vocabulary variation. In future work, we plan to explore larger dialectal datasets, improved contextual modeling techniques, and specialized methods for sarcasm detection to further enhance sentiment analysis for Arabic dialects.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT and MARBERT: Deep bidirectional transformers for arabic. In *Proceedings of ACL*.

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 Arabic NLP Workshop*.

Samhaa R. El-Beltagy, Mohamed El-Kalamawy, and Ali B. Soliman. 2017. Arabic sentiment analysis. In *SemEval*.

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 LREC 2026*, Palma, Mallorca, Spain.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *ICML*.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *SemEval*.