Dialect-Aware Sentiment Analysis for Ahasis Challenge

Hasna Chouikhi
LIMTIC Laboratory, University of Carthage,
Tunis, Tunisia
hasna.chouikhi@gmail.com

Manel ALOUI
SUP'COM, University of Carthage,
Tunis, Tunisia
manel.aloui@supcom.tn

Abstract

This paper presents our approach to Arabic sentiment analysis with a specific focus on dialect-awareness for Saudi and Moroccan (Darija) dialectal variants. We develop a system that achieves a macro F1 score of 77% on the test set, demonstrating effective generalization across these dialect variations. Our approach leverages a pre-trained Arabic language model (Qarib) with custom dialect-specific embeddings and preprocessing techniques tailored to each dialect. The results demonstrate a significant improvement over baseline models that do not incorporate dialect information, with an absolute gain of 5% in F1 score compared to the equivalent non-dialect-aware model. Our analysis further reveals distinct sentiment expression patterns between Saudi and Darija dialects, highlighting the importance of dialect-aware approaches for Arabic sentiment analysis.

1 Introduction

Sentiment analysis for Arabic text presents unique challenges due to the significant variations between Modern Standard Arabic (MSA) and regional dialects. These dialects differ in vocabulary, grammar, and expressions of sentiment, making cross-dialect sentiment analysis particularly challenging. This challenge is further compounded by the informal nature of social media text, where dialectal variations are prominent.

Our work focuses on developing a robust sentiment analysis system for Arabic social media reviews for hospitality that effectively handles dialectal variations, particularly between Saudi and Moroccan (Darija) dialects. We explore how dialect-aware modeling can improve sentiment classification accuracy and develop dialect-specific preprocessing techniques to normalize text while preserving sentiment information.

The variation in Arabic dialect poses significant challenges for NLP tasks due to the following:

- Lexical differences between dialects (different words for the same concept)
- Grammatical variations that affect sentence structure
- Cultural context and idiomatic expressions specific to each dialect
- Lack of standardized orthography for dialectal Arabic

Our approach addresses these challenges by combining dialect-specific preprocessing with a neural architecture that explicitly leverages dialect information during classification.

2 Related work

Recent advances in Arabic natural language processing have seen the development of several dialect-aware pre-trained language models. Models such as AraBERT (Antoun et al., 2020), MARBERT (Abdul-Mageed et al., 2021), CamelBERT (Inoue et al., 2021), and QARIB (Abdelali et al., 2021) have shown promising results for various Arabic NLP tasks. However, their effectiveness for dialect-specific sentiment analysis varies significantly.

Previous work on Arabic sentiment analysis has focused primarily on MSA or single-dialect approaches. (Al-Twairesh et al., 2017) explored sentiment analysis for the Saudi dialect, while (Oueslati et al., 2020) focused on the Tunisian dialect. Multi-dialect approaches, such as those presented in (Abdul-Mageed et al., 2012), have demonstrated that incorporating dialect information can improve performance; however, the optimal approach for dialect-aware sentiment analysis remains an open question.

Recent research has made significant strides in dialect-specific Arabic sentiment analysis through

Proceedings of the Shared Task on Sentiment Analysis for Arabic Dialects (AHaSIS) associated with RANLP 2025,
pages 40–45, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-109-7-007

novel datasets and advanced modeling techniques. (Hussein and Lakizadeh, 2024) introduced IRAQIDSAD, a benchmark dataset comprising 14,141 annotated comments in the Iraqi dialect, which addresses key challenges in dialectal Arabic syntax, morphology, and grammar. Their work includes a systematic review of the literature and corpus development methodology, providing a foundation for future research in Arabic sentiment analysis.

In a complementary effort, (BOUZIANE et al., 2024) demonstrated the effectiveness of Bi-LSTM networks for sentiment analysis on Algerian Arabic social media content, achieving state-of-the-art performance (94% accuracy). Their findings highlight the practical applications of such models in monitoring online discourse, guiding business strategies, and informing policy decisions.

Further advancing this domain, (Cherrat et al., 2024) explored the use of AraBERT and other deep learning approaches for sentiment analysis in the Moroccan dialect. Their results underscore the potential of transformer-based models to improve accuracy and generate nuanced insights into the opinions and emotions of Arabic-speaking populations.

Most prior work has treated Arabic dialects as separate languages, leading to the development of isolated models for each dialect. However, recent shared tasks such as **AHaSIS** have highlighted the importance of evaluating sentiment analysis in a variety of Arabic dialects using unified benchmarks and baselines (Alharbi et al., 2025a); (Alharbi et al., 2025b). In contrast to approaches that train separate models per dialect, our work proposes a unified model that processes multiple dialects simultaneously by explicitly incorporating dialectal identity as an input feature. This design enables the model to capture both shared patterns and dialect-specific nuances, improving generalization and performance in the analysis of the sentiment of Arabic dialects.

3 Dataset and Task Description

3.1 Dataset

The dataset consists of Arabic social media text predominantly from two dialects: Saudi and Moroccan (Darija). As shown in Fig. 2, the training set contains 860 samples, perfectly balanced between the two dialects (430 samples each). Each sample is annotated with one of three sentiment classes: positive, negative, or neutral.

The sentiment distribution in the training data, as illustrated in Fig. 1, shows: negative (336 samples), positive (308 samples) and neutral (216 samples), revealing a slight class imbalance that we address in our approach.

The test set contains 216 samples, also equally balanced between the two dialects (108 samples each). This balanced distribution allows an effective evaluation of the performance of the model in both dialects.

Figure 1: Distribution of sentiment classes in the training set

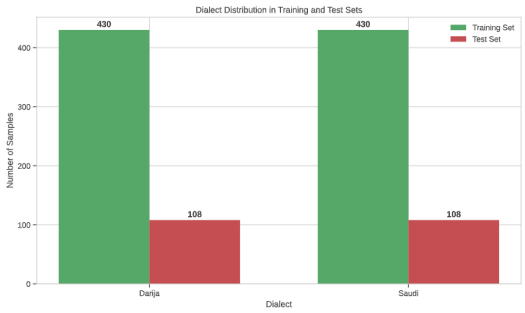

Figure 2: Distribution of dialects in training and test sets

3.2 Task Description

The task involves classifying the sentiment of Arabic text as positive, negative, or neutral, while effectively handling dialectal variations. Success is measured primarily by macro F1 score, with balanced accuracy as a secondary metric. Both metrics are important due to the class imbalance and the need to perform well across all sentiment categories.

4 Methodology

4.1 Model Architecture

Our model architecture is based on the pre-trained **Qarib** [1] model with significant customizations for

[1] https://huggingface.co/ahmedabdelali/bert-base-qarib

dialect-aware sentiment analysis. Fig. 3 provides a detailed illustration of our proposed architecture.

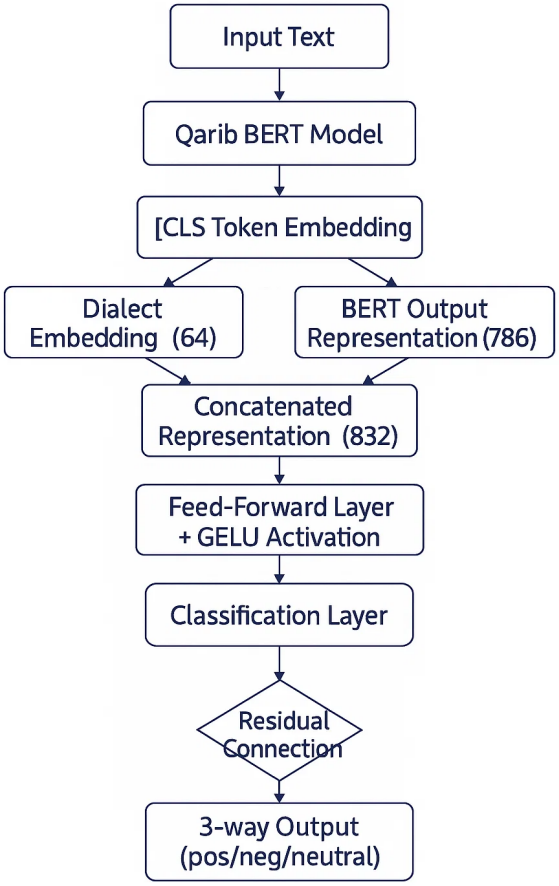

Figure 3: Dialect-Aware Sentiment Analysis Model Architecture

Key components of our architecture include:

a) **Dialect-Aware Embeddings**: We incorporate dialect information through specialized embeddings (64-dimensional) that are concatenated with the BERT base model output representation (768-dimensional) to create a combined 832-dimensional representation.

b) **Enhanced Classifier**: The classifier includes two feed-forward layers with GELU (Gaussian Error Linear Unit) activation function, layer normalization, and residual connections. This design helps the model better capture the complex relationship between dialect-specific features and sentiment expressions.
Mathematically, the GELU activation function is defined as:

$$\text{GELU}(x) = x \cdot \Phi(x) \quad (1)$$

where $\Phi(x)$ is the cumulative distribution function (CDF) of the standard normal distribution.

An efficient approximation, commonly used in practice, is given by:

$$\text{GELU}(x) \approx 0.5x \left(1 + \tanh\left[\sqrt{\tfrac{2}{\pi}}\left(x + 0.044715x^3\right)\right]\right) \quad (2)$$

This approximation provides a smooth, nonlinear transformation that retains the stochastic regularization properties of GELU while being computationally more efficient.

c) **Dialect-Specific Preprocessing**: We implement custom preprocessing for Saudi and Darija dialects, normalizing dialectal variations while preserving sentiment indicators.

4.2 Preprocessing

We developed dialect-specific preprocessing techniques to normalize the text while preserving dialect-specific sentiment markers:

a) **General Arabic Normalization**:

- Removing diacritics (tashkeel)
- Normalizing various forms of alef (إ أ آ ا → ا)
- Normalizing hamzas (ؤ ئ → ء)
- Normalizing yaa and taa marbuta (ى → ي, ة → ه)
- Removing tatweel/kashida (ـ)

b) **Dialect-Specific Normalization**:

- For Saudi dialect, we normalize common expressions like "مره" → "مرة" ("very" or "really"), "كمان" → "أيضا" ("also" or "too"), etc.
- For Darija, we normalize expressions like "بزاف" → "كثير" ("a lot" or "very much"), "ماشي" → "ليس" ("not" or "no"), etc.

This preprocessing strategy helps standardize the input while retaining crucial dialect-specific sentiment indicators, creating a more consistent representation for the model.

4.3 Training Methodology

Our training approach incorporates several techniques to address the challenges of dialectal sentiment analysis:

a) **Focal Loss**: We use focal loss with gamma=2.0 to address class imbalance and focus on hard examples.

b) **Class Weighting**: We apply balanced class weights to address the imbalance between sentiment classes, particularly the underrepresented neutral class.

c) **Learning Rate Schedule**: We employ a linear warmup followed by linear decay, with a maximum learning rate of 2e-6.

d) **Gradient Accumulation**: We use 4 gradient accumulation steps to achieve an effective batch size of 64 while maintaining memory efficiency.

e) **Discriminative Fine-tuning**: We apply different learning rates across model layers, with lower rates for embeddings and early layers, and higher rates for task-specific layers.

4.4 Hyperparameters

Our final model uses the following hyperparameters:

- Model: ahmedabdelali/bert-base-qarib
- Maximum Sequence Length: 128
- Batch Size: 16 (Effective Batch Size: 64 with gradient accumulation)
- Learning Rate: 2e-6
- Epochs: 10
- Early Stopping Patience: 3
- Focal Loss Gamma: 2.0
- Dialect Embedding Size: 64
- Scheduler: Linear with Warmup (15%)
- Dropout Rate: 0.2

5 Experimental Results

5.1 Overall Performance

Our best model achieved the following results on the test set:

- Macro F1 Score: 0.770
- Balanced Accuracy: 0.775
- Precision: 0.771
- Recall: 0.769

As shown in Fig. 4, our dialect-aware approach significantly outperforms baseline models and non-dialect-aware variants:

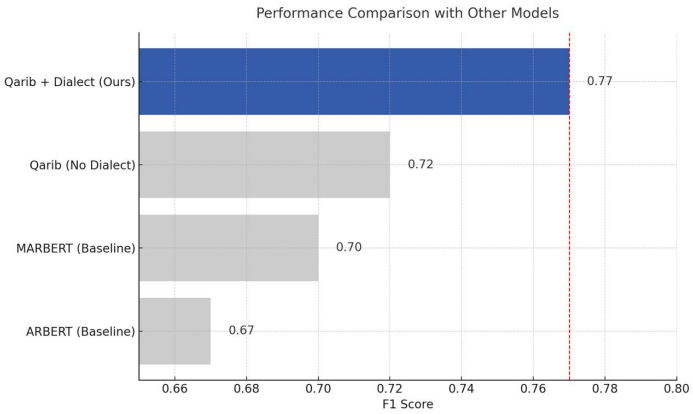

Figure 4: Performance comparison with other models

The results demonstrate that our dialect-aware approach achieves a 5% absolute improvement in F1 score compared to the same model without dialect features (0.77 vs. 0.72), and even larger improvements over general Arabic language models MARBERT (0.70) and ARBERT (0.67).

5.2 Performance by Dialect

The model showed different patterns across dialects:

Table 1: Performance Metrics by Dialect

Dialect	F1 Score	Accuracy	Precision	Recall
Saudi	0.787	0.787	0.792	0.787
Darija	0.752	0.752	0.750	0.752

These results indicate that the model performs better on Saudi dialect than on Darija, though the performance is strong for both dialects. This difference may be attributed to the inherent complexity of Darija, which incorporates influences from Berber, French, and Spanish.

5.3 Performance by Sentiment Class

Table 2: Performance Metrics by Sentiment Class

Sentiment	Precision	Recall	F1 Score
Positive	0.810	0.765	0.787
Negative	0.825	0.743	0.782
Neutral	0.678	0.800	0.734

These results show that the model performs best on positive and negative sentiment detection, while neutral sentiment is more challenging. The neutral class has the lowest precision but highest recall, indicating some tendency to classify ambiguous cases as neutral.

6 Discussion and Analysis

6.1 Dialect-Specific Patterns

Our analysis revealed distinct sentiment expression patterns between Saudi and Darija dialects:

a) **Saudi Dialect**:

- More direct expressions of sentiment
- Higher proportion of positive sentiment
- Lower use of neutral expressions
- Cultural references specific to Gulf regions

b) **Darija Dialect**:

- More circumspect sentiment expressions
- Higher proportion of neutral statements
- Context-dependent interpretation more common
- Borrowings from French and Berber that affect sentiment expression

These patterns highlight the importance of dialect-specific approaches to sentiment analysis in Arabic. For example, certain expressions in Saudi dialect are inherently positive or negative, while similar constructions in Darija might be more neutral or ambiguous without additional context.

6.2 Error Analysis

Analysis of misclassifications revealed several patterns, including the need for manual verification of annotations, as some sentences labeled as neutral contained implicit positive or negative sentiment. Other key challenges included:

a) **Sarcasm and Irony**: The model struggled with sarcastic expressions, particularly in Darija dialect where sarcasm is often marked by subtle contextual cues rather than explicit markers.

b) **Context-Dependent Sentiment**: Cases where sentiment depended on broader cultural or situational context were challenging, as the model lacked access to this external information.

c) **Dialect Misidentification**: Some errors stemmed from incorrect dialect identification, particularly for less distinctive dialect markers.

These findings suggest that improving annotation quality—especially for implicitly subjective text—along with better handling of sarcasm, context, and multilingualism, could further enhance model performance.

6.3 Impact of Dialect-Aware Features

The dialect-specific embeddings proved crucial for performance, improving F1 score by 5% absolute. This confirms our hypothesis that dialect information is essential for accurate sentiment analysis in dialectal Arabic.

The improvement was particularly pronounced for the neutral class, where dialect awareness helped distinguish between genuinely neutral statements and culturally-specific expressions that might appear neutral without dialect context.

7 Conclusion and Future Work

This paper presented a dialect-aware sentiment analysis approach for Arabic social media text that performs well across Saudi and Darija dialects. Our model effectively incorporates dialect information through specialized embeddings and preprocessing, demonstrating the importance of dialect awareness for Arabic sentiment analysis.

Key findings include the significant improvement in sentiment classification when using dialect-specific features, with our model achieving 77% F1 score compared to 72% without dialect features. Additionally, we observed that different dialects exhibit distinct patterns in sentiment expression, highlighting the need for tailored approaches. Furthermore, we showed that class imbalance can be effectively addressed through focal loss and class weighting techniques.

For future work, several directions could be explored. First, expanding the model to additional Arabic dialects would enhance its generalizability. Second, incorporating external knowledge sources could help capture culturally specific expressions more accurately. Third, exploring multi-task learning with explicit dialect identification might further improve performance. Finally, addressing code-switching through multilingual approaches could make the model more robust in real-world scenarios where users mix languages and dialects.

References

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training bert on arabic tweets: Practical considerations.

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 (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Muhammad Abdul-Mageed, Sandra Kuebler, and Mona Diab. 2012. SAMAR: A system for subjectivity and sentiment analysis of Arabic social media. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 19–28, Jeju, Korea. Association for Computational Linguistics.

Nora Al-Twairesh, Hend Al-Khalifa, AbdulMalik Al-Salman, and Yousef Al-Ohali. 2017. Arasenti-tweet: A corpus for arabic sentiment analysis of saudi tweets. *Procedia Computer Science*, 117:63–72. Arabic Computational Linguistics.

Maram Alharbi, Salmane Chafik, Saad Ezzini, Ruslan Mitkov, Tharindu Ranasinghe, and Hansi Hettiarachchi. 2025a. Ahasis: Shared task on sentiment analysis for arabic dialects. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*.

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

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, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Abdelghani BOUZIANE, Benamar BOUOUGADA, Djelloul BOUCHIHA, and Noureddine DOUMI. 2024. Sentiment analysis of algerian arabic dialect on social media using bi-lstm recurrent neural networks. *The Journal of Engineering and Exact Sciences*, 10(7):20058–20058.

El Mehdi Cherrat, Hassan Ouahi, Abdellatif BEKKAR, et al. 2024. Sentiment analysis from texts written in standard arabic and moroccan dialect based on deep learning approaches. *International Journal of Computing and Digital Systems*, 16(1):447–458.

Hafedh Hameed Hussein and Amir Lakizadeh. 2024. Iraqidsad: A dataset for benchmarking sentiment analysis tasks on iraqi dialect based texts. *International Journal of Advances in Soft Computing & Its Applications*, 16(3).

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.

Oumaima Oueslati, Erik Cambria, Moez Ben HajHmida, and Habib Ounelli. 2020. A review of sentiment analysis research in arabic language. *Future Generation Computer Systems*, 112:408–430.