# Sentiment Analysis on Arabic Dialects: A Multi-Dialect Benchmark

**Abdusalam F A Nwesri** University of Tripoli a.nwesri@uot.edu.ly

**Nabila Almabrouk S. Shinber** College of Science and Technology shinbir@tcst.edu.ly

**Amani Bahlul Sharif** University of Tripoli am.sharif@uot.edu.ly

## Abstract

This paper presents our contribution to the AHASIS Shared Task at RANLP 2025, which focuses on sentiment analysis for Arabic dialects. While sentiment analysis has seen considerable progress in Modern Standard Arabic (MSA), the diversity and complexity of Arabic dialects pose unique challenges that remain underexplored. We address this by fine-tuning six pre-trained language models, including AraBERT, MARBERTv2, QARiB, and DarijaBERT, on a sentiment-labeled dataset comprising hotel reviews written in Saudi and Moroccan (Darija) dialects. Our experiments evaluate the models' performance on both combined and individual dialect datasets. MARBERTv2 achieved the highest performance with an F1-score of 79% on the test set, securing third place among 14 participants. We further analyze the effectiveness of each model across dialects, demonstrating the importance of dialect-aware pre-training for Arabic sentiment analysis. Our findings highlight the value of leveraging large pre-trained models tailored to dialectal Arabic for improved sentiment classification.

## 1 Introduction

Arabic is the official language of 22 countries. It comprises of Modern Standard Arabic (MSA) which used for formal writing and a wide array of spoken dialects. Dialects have been used purely for communication as a spoken version of Arabic. However, with the domination of the social media, dialects have transformed into written text format. Huge text is written in local dialects describing different opinions, emotions and personal thoughts.

Dialects differ significantly from MSA and from each other in syntax, phonology, morphology, vocabulary, and orthography(Habash et al., 2018). New NLP techniques and approaches are required to understand each individual dialect. In this work, we explore using pre-trained Large Language Models (LLM) in sentiment analysis for text written in two dialects. The work was done at the AHASIS shared task organized by the RANLP 2025 (Alharbi et al., 2025b).

The following sections detail the shared task, the dataset employed, and the conducted experiments along with their corresponding results.

## 2 Related work

Several studies have been conducted in the field of Arabic sentiment analysis, primarily focusing on texts written in MSA, whereas Arabic dialects have remained relatively underexplored (Shi and Agrawal, 2025; Boudad et al., 2018).

### 2.1 Lexicon-based approaches

Early work in Arabic sentiment analysis primarily relied on lexicon-based approaches (Badaro et al., 2014; Al-Moslmi et al., 2017). In this approach a specialized lexicon is constructed where words are annotated with polarity or sentiment scores. This lexicon is then used to calculate the whole text sentiment by summing up sentiment scores of its words. Most popular sentiment lexicons are ArsenL (Arabic Sentiment Lexicon)(Badaro et al., 2014) and Arabic Senti-Lexicon (Al-Moslmi et al., 2017).

### 2.2 Machine Learning approaches

The lexicon-based methods need human effort and word scores are not accurate as words usually appear in different context. To overcome these limitations, Machine Learning (ML) techniques were used. In such techniques, word scores are calculated using feature engineering such as the bag-of-words(Qader et al., 2019), TF-IDF (Sammut and Webb, 2010) and word embedding (Almeida and Xexéo, 2019). Then ML algorithms such as Naive

*Proceedings of the Shared Task on Sentiment Analysis for Arabic Dialects (AHaSIS) associated with RANLP 2025*, pages 86–91, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-109-7-013

Bayes and Support Vector Machines (SVMs) are used to identify text sentiment.

### 2.3 Large Language Models

The success of pre-trained language models based on bidirectional transformers—such as BERT(Devlin et al., 2019) across various natural language understanding tasks has led to growing interest in their application to Arabic sentiment classification. ElJundi et al. (2019) introduced hULMonA, a language model tailored specifically for Arabic, which they fine-tuned for sentiment analysis. The model achieved 95% on F1 when using the Hotel Arabic Reviews Dataset (HARD) (Elnagar et al., 2018); a combination of MSA and Golf dialect text. However, the model result was only 50% when using the Arabic Sentiment Twitter Dataset for LEVantine dialect (ArSenTD-Lev) dataset (Baly et al., 2019).

Abdul-Mageed et al. (2021) presented ARBERT and MARBERT models. The ARBERT model is trained on MSA data, and as the authors mentioned, is not best suited for downstream tasks involving dialectal Arabic. For such tasks, they introduced the MARBERT, a model which is trained on a large Twitter dataset that includes Arabic dialects text. MARBERT was reported to be superior to most of the state-of-the-art models in several tasks specially when using social media datasets. The model has been used afterwards in several tasks including dialect identification tasks(Nwesri et al., 2023), offensive language detection(Abdellaoui et al., 2024).

Another model which is trained on both MSA and dialects and has been reported to perform well on sentiment analysis is the QCRI Arabic and Dialectal BERT (QARiB) model, the model was trained on a collection of 420 million tweets and about 180 Million sentences of text. It was reported that QARiB achieved an accuracy of 93% on sentiment analysis task involving Darija text (Bourahouat et al., 2023).

Some specific dialect BERT-based models have been introduced. DarijaBERT, SudaBERT, TunBERT, and DziriBERT are models trained on Moroccan, Sudanise, Tunisian, and Algerian dialects respectively. The DarijaBERT was reported to be effective in a sentiment analysis task using Maghribi Dialect. The model scores 92% on F1-score.

In this study, we focus on fine-tuning models that were pre-trained on multi-dialectal text and have been reported to perform well on the task of Arabic dialect sentiment analysis.

## 3 Methodology

### 3.1 Task

The shared task is organized as part of RANLP 2025. The task purpose is to advance the Arabic dialectal sentiment analysis and generate a benchmark for a Multi-Dialect sentiment analysis on hotel reviews (Alharbi et al., 2025a).

Participants are provided with a multi-dialectal annotated dataset and engage in sentiment detection in Arabic dialects. The task aim is to address the challenges of dialect-specific sentiment detection, cross-lingual sentiment preservation, and nuanced sentiment classification in customer reviews of hotels.

### 3.2 Dataset

The dataset consists of 860 hotel reviews written in Saudi and Maghribi (Darija) dialects. Text reviews are annotated as either “Positive”, “Negative” or “Neutral”. The dialect is also included along with each review. Both Saudi and Darija subsets contain 154 positive, 168 negative, and 108 neutral reviews.

The test set, released later during the evaluation phase, comprises 216 reviews—108 each in Saudi and Darija dialects. Participants are required to predict the sentiment polarity of these reviews using their developed models. An additional column containing the predicted labels must be appended to the test set prior to submission on the shared task platform, where automatic evaluation is conducted. Each participant is allowed a maximum of five submissions during the evaluation phase. The results are displayed on a public leaderboard, showcasing the performance of all participating teams.

### 3.3 Models

The baseline model for the task was the Pre-trained BERT-based model (AraBERT) fine-tuned on MSA and Arabic dialect data (Antoun et al., 2021). We have focused on fine-tuning the State-of-the-art models which have been trained on both MSA and Arabic dialects. Basically, we fine-tuned the bert-base-arabert, bert-base-arabertv02-twitter, bert-large-arabertv02-twitter, MARBERTv2, bert-base-qarib, and DarijaBERT models.

| hyperparameter | From | To |
|---|---|---|
| learning rate | 1e-5 | 2e-2 |
| Training batch size | 8 | 64 |
| Evaluation batch size | 8 | 32 |
| weight decay | 0.1 | 0.3 |
| warm-up ratio | 1e-4 | 0.1 |
| number of epochs | 4 | 10 |

Table 1: Summary of hyperparameter ranges used in our experiments

### 3.4 Evaluation Measure

The organizers used the F1-score as the primary metric to evaluate the performance of various models. Additionally, sentiment accuracy comparisons across dialects was also considered.

### 3.5 Baseline System

The pre-trained BERT-based model (AraBERT) fine-tuned on MSA and Arabic dialect data was set as the baseline system by the organizers. Participants were encouraged to improve upon the baseline model with their own techniques and use LLMs.

## 4 Experiments

Several Arabic pre-trained models have been fine-tuned on this task. The choice of hyper-parameters ranges considerably between models. Table 1 shows the ranges we used in our experiments. They are learning rate, batch size, weight decay, warm-up steps, and the number of epochs. We chose epochs as the evaluation strategy and used the F1-score as the metrics for the evaluation.

We fine-tuned six pre-trained models namely: bert-base-arabert, bert-base-arabertv02-twitter, bert-large-arabertv02-twitter, MARBERTv2, bert-base-qarib, and DarijaBERT models using HuggingFace's Trainer API. All experiments are run using the google colab platform to run our experiments (Bisong, 2019).

In all experiments, the training dataset was split into 80% for training and 20% for evaluation. Truncation and padding were applied with a maximum sequence length of 128 tokens during preprocessing using the model's tokenizer.

### 4.1 Experiment 1: Fine-tuning using the original training dataset

In this experiment, we fine-tuned the models using the original dataset. Table 2 presents the best results achieved by each model. All models demonstrated strong performance on the training data, with their optimal results determined through careful hyperparameter tuning. MARBERTv2 outperformed the others, achieving an F1-score of 88%, followed by bert-large-arabert and bert-base-qarib.

Table 3 presents the official results submitted to the shared task website. The best-performing model on the test set was MARBERTv2, achieving an F1-score of 79%. bert-base-qarib followed with 77%, and DarijaBERT scored 76%. Due to submission limits during the evaluation phase, not all results were submitted. The 79% score earned us third place among the 14 participating teams.

### 4.2 Experiment 2: Fine-tuning using a single dialect

In this experiment, we evaluate the performance of selected models in sentiment detection using text from a single dialect. To do so, the training dataset was divided into two subsets: one containing 430 reviews in the Saudi dialect and another with 430 reviews in the Moroccan (Darija) dialect. All models were tested on both datasets using identical hyperparameter settings. basically lr is set to 1e-4, both training and evaluation batch sizes are set to 8, number of epoch is 4, weight decay is set to 0.01, and warmup ratio is set to 0.3. Table 4 Shows the models performance on both datasets.

The best-performing model on the Saudi dialect is bert-large-arabertv02-twitter, achieving an F1-score of 85%, followed by bert-base-arabertv02-twitter and MARBERTv2. For the Darija dialect, MARBERTv2 led with a 78% F1-score, followed by bert-large-arabertv02-twitter and bert-base-arabertv02-twitter. The bert-base-qarib model showed consistent performance across both dialects, while DarijaBERT surprisingly performed better on the Saudi dialect.

### 4.3 Experiment 3: Fine-tuning using a multi-dialect text

To investigate the impact of text length and the presence of multiple dialects on sentiment analysis, we created a mixed-dialect dataset by combining tweets written in Saudi and Darija dialects. Using matching review IDs from the dataset, we merged corresponding reviews from both dialects to form a single entry containing text from both dialects along with its sentiment label. The models were then used to predict sentiment on this mixed dataset. To accommodate the longer input, the maximum

|  | Hyperparameters |  |  |  |  |  | Result |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | lr | tbs | ebs | ep | wd | wr | Acc. | F1 | P | R |
| bert-base-arabert | 7e-5 | 32 | 32 | 10 | 1e-4 | 0.3 | 0.854 | 0.854 | 0.854 | 0.854 |
| bert-base-arabertv02-twitter | 1e-4 | 16 | 16 | 4 | 1e-4 | 0.1 | 0.872 | 0.872 | 0.872 | 0.872 |
| bert-large-arabertv02-twitter | 1e-4 | 8 | 8 | 3 | 1e-4 | 0.1 | 0.878 | 0.878 | 0.878 | 0.878 |
| MARBERTv2 | 2e-4 | 16 | 16 | 4 | 1e-4 | 0.1 | **0.884** | **0.884** | **0.884** | **0.884** |
| bert-base-qarib | 2e-4 | 16 | 16 | 5 | 1e-4 | 0.1 | 0.877 | 0.861 | 0.872 | 0.859 |
| DarijaBERT | 2e-4 | 16 | 8 | 5 | 1e-4 | 0.1 | 0.849 | 0.849 | 0.849 | 0.849 |

Table 2: Training results. lr=learning rate, tbs=training batch size, ebs= Evaluation batch size, wr=warmup ratio, wd= weight decay, and ep=number of epochs.

| Model | Accuracy | F1-score | Precision | Recall | Balanced Accuracy |
|---|---|---|---|---|---|
| MARBERTv2 | 0.792 | 0.792 | 0.792 | 0.792 | 0.784 |
| bert-base-qarib | 0.773 | 0.773 | 0.773 | 0.773 | 0.765 |
| DarijaBERT | 0.759 | 0.759 | 0.759 | 0.759 | 0.758 |

Table 3: Official submitted runs.

sequence length for all models was increased to 512 tokens. Table 5 presents the performance of the six models on this dataset.

The bert-large-arabertv02-twitter model achieved the highest performance, scoring 84% across all metrics, followed by bert-base-arabertv02-twitter with 83% and MARBERTv2 with 81%. bert-base-qarib and bert-base-arabert delivered comparable results, with F1-scores of 80% and 79%, respectively. The DarijaBERT had the lowest performance among the models, with scoring 73% on all metrics. Overall, the Arabert-based models, particularly the v02-twitter variant, demonstrated superior effectiveness on this experiment.

## 5 Discussion

Our experiments reveal several key insights into the behavior of modern pre-trained Arabic LLMs when applied to dialectal sentiment analysis. First, across all settings, models that have been pre-trained on large, diverse social-media corpora (i.e. the Arabertv02-twitter variants and MARBERTv2) consistently outperform both the smaller AraBERT base model and the dialect-specific DarijaBERT. This suggests that broad exposure to multiple dialects and informal text during pre-training is more beneficial than narrow, single-dialect pre-training, even when the downstream data are from just one dialect.

Second, the relative ranking of models remains largely stable across our three experimental scenarios (original mixed-dialect training set, single-dialect subsets, and mixed long-review set). In particular, MARBERTv2, bert-large-arabertv02-twitter and bert-base-arabertv02-twitter occupy the top three positions in almost every setting. This robustness indicates that careful hyperparameter tuning and increased model capacity (i.e. “large” vs. “base”) yield consistent gains even as input characteristics (dialect, length, mixing) vary.

Third, fine-tuning on single-dialect subsets highlights subtle dialectal biases: MARBERTv2 performs best on Darija, while bert-large-arabertv02-twitter excels on the Saudi subset. This confirms that some models encode dialect-specific patterns more strongly, likely reflecting the composition of their pre-training data. Yet, the performance drop when moving from single-dialect to mixed long reviews is relatively modest (only 2–3%), indicating that these models can generalize to code-switched or concatenated dialect inputs—an encouraging result for real-world social-media applications.

Finally, DarijaBERT’s lower scores (around 73%) across all settings underscore the limitations of highly specialized dialect BERTs when applied outside of their narrow target domain or when compared to larger, multi-dialectal LLMs. Future work should explore whether further increases in data volume, additional dialects, or adapter-based approaches can close this gap.

## 6 Conclusion

In this work, we fine-tuned six state-of-the-art Arabic pre-trained models on the RANLP 2025 AHASIS shared task dataset to evaluate their effective-

| | Saudi Dialect | | | | Darija Dialect | | | |
|---|---|---|---|---|---|---|---|---|
| Model | Acc. | F1 | P | R | Acc. | F1 | P | R |
| bert-base-arabert | 0.733 | 0.733 | 0.733 | 0.733 | 0.709 | 0.709 | 0.709 | 0.709 |
| bert-base-arabertv02-twitter | 0.814 | 0.814 | 0.814 | 0.814 | 0.767 | 0.767 | 0.767 | 0.767 |
| bert-large-arabertv02-twitter | **0.849** | 0.849 | 0.849 | 0.849 | 0.767 | 0.767 | 0.767 | 0.767 |
| MARBERTv2 | 0.802 | 0.802 | 0.802 | 0.802 | **0.779** | 0.779 | 0.779 | 0.779 |
| bert-base-qarib | 0.733 | 0.733 | 0.733 | 0.733 | 0.733 | 0.733 | 0.733 | 0.733 |
| DarijaBERT | 0.733 | 0.733 | 0.733 | 0.733 | 0.698 | 0.698 | 0.698 | 0.698 |

Table 4: Results of fine-tuning models on the Saudi and Moroccan dialects.

| Model | Acc. | F1 |
|---|---|---|
| bert-base-arabert | 0.790 | 0.790 |
| bert-base-arabertv02-twitter | **0.837** | **0.837** |
| bert-large-arabertv02-twitter | 0.826 | 0.826 |
| MARBERTv2 | 0.814 | 0.814 |
| bert-base-qarib | 0.802 | 0.802 |
| DarijaBERT | 0.732 | 0.732 |

Table 5: Results of fine-tuning models on the mixed long reviews dataset.Recall and Precision columns have the same values across all models.

ness on multi-dialect hotel review sentiment analysis. Our key findings are:

- Models pre-trained on large, multi-dialect Twitter corpora (Arabertv02-twitter and MARBERTv2) consistently outperform both standard AraBERT and dialect-specific BERTs.

- Increasing model capacity (large vs. base) and careful hyperparameter tuning yield reliable performance improvements across varied input scenarios.

- While certain models exhibit dialectal biases (e.g. MARBERTv2 on Darija, bert-large-arabertv02-twitter on Saudi), all top models maintain high accuracy and F1 ( $>$ 81%) even on mixed, longer inputs.

- Narrowly trained dialect BERTs (DarijaBERT) lag behind, highlighting the value of broad, multi-dialectal pre-training.

Our best submission, MARBERTv2, achieved 79% F1 on the blind test set, ranking third among 14 teams. Future directions include exploring adapter-based fine-tuning to reduce resource demands, incorporating explicit dialect identifiers, and extending experiments to additional dialects and domains to further assess model generalizability.

## References

Israe Abdellaoui, Anass Ibrahimi, Mohamed Amine El Bouni, Asmaa Mourhir, Saad Driouech, and Mohamed Aghzal. 2024. Investigating offensive language detection in a low-resource setting with a robustness perspective. *Big Data and Cognitive Computing*, 8(12).

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.

Tareq Al-Moslmi, Mohammed Albared, Adel Al-Shabi, Nazlia Omar, and Salwani Abdullah. 2017. Arabic senti-lexicon: Constructing publicly available language resources for arabic sentiment analysis. *Journal of Information Science*, 44:345–362.

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)*.

Felipe Almeida and Geraldo Xexéo. 2019. Word embeddings: A survey. *CoRR*, abs/1901.09069.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. Arabert: Transformer-based model for arabic language understanding.

Gilbert Badaro, Ramy Baly, Hazem Hajj, Nizar Habash, and Wassim El-Hajj. 2014. A large scale Arabic sentiment lexicon for Arabic opinion mining. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 165–173, Doha, Qatar. Association for Computational Linguistics.

Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. 2019. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2109–2116, Minneapolis, Minnesota. Association for Computational Linguistics.

Ekaba Bisong. 2019. *Google Colaboratory*, pages 59–64. Apress, Berkeley, CA.

Naaima Boudad, Rdouan Faizi, Rachid Oulad Haj Thami, and Raddouane Chiheb. 2018. Sentiment analysis in arabic: A review of the literature. *Ain Shams Engineering Journal*, 9(4):2479–2490.

Ghizlane Bourahouat, Manar Abourezq, and Najima Daoudi. 2023. Leveraging moroccan arabic sentiment analysis using arabert and qarib. In *Innovations in Smart Cities Applications Volume 6*, pages 299–310, Cham. Springer International Publishing.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Obeida ElJundi, Wissam Antoun, Nour El Droubi, Hazem Hajj, Wassim El-Hajj, and Khaled Shaban. 2019. hULMonA: The universal language model in Arabic. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 68–77, Florence, Italy. Association for Computational Linguistics.

Ashraf Elnagar, Yasmin S. Khalifa, and Anas Einea. 2018. *Hotel Arabic-Reviews Dataset Construction for Sentiment Analysis Applications*, pages 35–52. Springer International Publishing, Cham.

Nizar Habash, Fadhl Eryani, Salam Khalifa, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghouani, Houda Bouamor, Nasser Zalmout, Sara Hassan, Faisal Al-Shargi, Sakhar Alkhereyf, Basma Abdulkareem, Ramy Eskander, Mohammad Salameh, and Hind Saddiki. 2018. Unified guidelines and resources for Arabic dialect orthography. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Abduslam F A Nwesri, Nabila A S Shinbir, and Hassan Ebrahem. 2023. UoT at NADI 2023 shared task: Automatic Arabic dialect identification is made possible. In *Proceedings of ArabicNLP 2023*, pages 620–624, Singapore (Hybrid). Association for Computational Linguistics.

Raheel Qader, François Portet, and Cyril Labbé. 2019. Semi-supervised neural text generation by joint learning of natural language generation and natural language understanding models. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 552–562, Tokyo, Japan. Association for Computational Linguistics.

Claude Sammut and Geoffrey I. Webb, editors. 2010. *Encyclopedia of Machine Learning*. Springer.

Zhiqiang Shi and Ruchit Agrawal. 2025. A comprehensive survey of contemporary Arabic sentiment analysis: Methods, challenges, and future directions. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3760–3772, Albuquerque, New Mexico. Association for Computational Linguistics.