Lab17 @ Ahasis Shared Task 2025: Fine-Tuning and Prompting techniques for Sentiment Analysis of Saudi and Darija Dialects

Al Mukhtar Al Hadrami | Firas Al Mahrouqi | Mohammed Al Shaail | Hala Mulki

Sultan Qaboos University, Computer Science / Muscat, Oman

{132962, 133570, 134365 }@student.squ.edu.om

h.mulki@squ.edu.om

Abstract

In this paper, we describe our contribution to the Ahasis shared task on sentiment analysis of Arabic dialects in the hospitality domain. As part of this task, we evaluated two well-established learning strategies using a Large Language Model (LLM) and Transformer-based model variants. Specifically, we applied few-shot prompting with GPT-4o and conducted fine-tuning experiments on two models: the original MARBERT model using the official Ahasis dataset, and SODA-BERT, a MARBERT variant previously fine-tuned on an Omani sentiment dataset. Our results showed that few-shot prompting with GPT-4o achieved an F1 score of 69%. However, both MARBERT and SODA-BERT outperformed GPT-4o when fine-tuned on relevant data. In the official ranking, our system based on fine-tuned MARBERT achieved 8th place among the participating teams.

1 Introduction

Sentiment analysis—also known as opinion mining—is the automatic processing of text to identify and categorize the author's attitude or emotional tone (positive, negative, neutral), enabling large-scale insights into public opinion (Pang & Lee, 2008).

Arabic sentiment analysis is especially challenging due to its rich morphology, pervasive diglossia (Modern Standard Arabic vs. dialects), orthographic variation, and the scarcity of resources for many dialects, most notably Saudi and Moroccan Darija, which remain under-represented in existing corpora (Habash, 2010).

While a growing number of models target Egyptian or Levantine sentiment, far fewer have been trained or evaluated on Saudi or Darija data. This imbalance leaves a critical gap: the absence of large, high-quality annotated datasets for these under-represented dialects limits model generalization and performance (Zahran & Elglaly, 2024)

In this context, we present the Lab17 system, developed as a baseline submission to the Ahasis (Ahasis Shared Task, 2025) shared task on sentiment analysis in the hospitality domain. We aim to evaluate the effectiveness of established modeling techniques in low-resource dialectal settings, rather than proposing novel algorithms.

Our methodology explores three complementary strategies: few-shot prompting using GPT-4o, evaluation of SODA-MARBERT (a MARBERT-derived model fine-tuned on an Omani-dialect sentiment dataset), and direct fine-tuning of the MARBERT model on the official Ahasis training set.

Our experiments showed that, few-shot prompting with GPT-4o achieved an F1 score of 69%, SODA-MABERT reached 71%, and fine-

Proceedings of the Shared Task on Sentiment Analysis for Arabic Dialects (AHaSIS) associated with RANLP 2025, pages 35–39, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-109-7-006

tuning MARBERT on the shared-task data yielded the best result with an F1 score of 75%.

Following the shared task's official evaluation, Lab17 placed 8th overall, confirming the value of dialect-aware preprocessing and domain-specific fine-tuning for sentiment analysis in under-represented Arabic dialects.

2 Arabic Sentiment Analysis Approaches

Arabic Sentiment analysis has significantly evolved through four main paradigms: (1) traditional machine-learning classifiers, (2) pretrained embedding–based deep models, (3) transformer-based architectures, and (4) generative language models. Here, we briefly review each category, highlighting recent benchmark results.

2.1 Traditional Machine-Learning Classifiers

Support-Vector Machines (SVMs) and Random Forests remain robust baselines for dialectal and MSA sentiment tasks. Abdelwahab et al. (2022) applied an SVM with TF-IDF features to classify Egyptian-dialect tweets, achieving an F_1-score of 85.6 % (Abdelwahab et al., 2022). Likewise, Alsayat and El-Sayed (2022) demonstrated that a Random Forest classifier attained an F_1-score of 84.2 % on Saudi-dialect Twitter data (Alsayat & El-Sayed, 2022).

2.2 Pretrained Embedding Models

Distributed embeddings automate feature extraction and boost downstream classifiers. Al-Twairesh et al. (2023) utilized Word2Vec embeddings customized for Jordanian-dialect tweets, significantly improving sentiment-classification accuracy over handcrafted features (Al-Twairesh et al., 2023). Similarly, Alsarhan and Bouamor (2023) showed that Sentence2Vec embeddings raised the F_1-score for Gulf-dialect sentiment analysis by several points compared to standard Word2Vec (Alsarhan & Bouamor, 2023).

2.3 Transformer-Based Architectures

Self-attention models have set new state-of-the-art results. AraBERT, introduced by Antoun et al. (2020), fine tuning is the process of re-training a pre-trained model on new data, when fine-tuned on mixed MSA and dialectal tweets, consistently exceeds 90 % F_1 across multiple benchmarks (Antoun et al., 2020). Building on this, MARBERT was designed for dialectal Arabic; Abdel-Salam and Mubarak (2023) report that MARBERT achieves a 92.1 % F_1-score on diverse dialectal sentiment datasets (Abdel-Salam & Mubarak, 2023).

2.4 Generative Language Models

GPT-style models excel at zero- and few-shot sentiment tasks without extensive retraining , zero-shot learning refers to a model's ability to perform tasks without any prior examples during training, while few-shot learning enables the model to generalize from only a small number of examples. Mubarak et al. (2023) introduced AraGPT and demonstrated robust zero-shot sentiment classification on Gulf-dialect tweets, achieving an F_1 of 88.4 % (Mubarak et al., 2023). In follow-up work, Al-Khamissi et al. (2023) evaluated AraGPT2 on Emirati Instagram comments and confirmed its strong performance, reporting an 84.1 % F_1-score (Al-Khamissi et al., 2023).

3 Lab17 Shared Task Baseline:

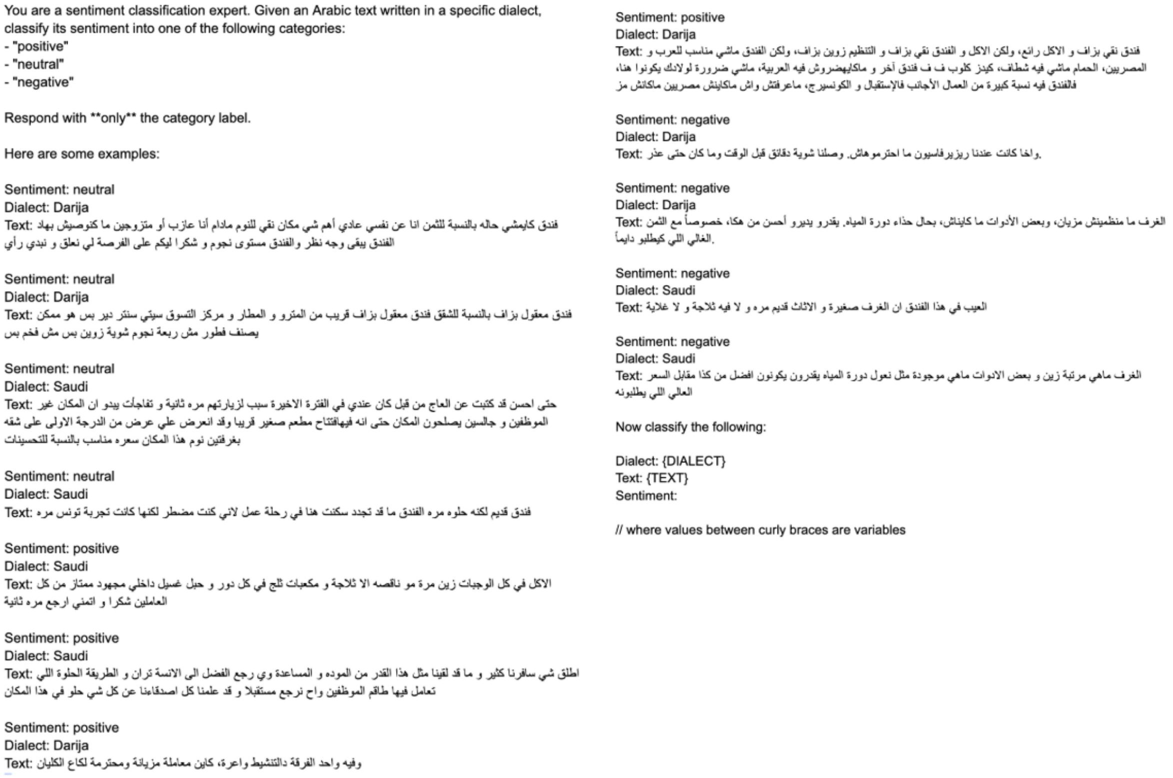
As part of our shared task submission, we implemented and compared three standard strategies for classifying the sentiment of tweets written in Saudi and Darija dialects. The objective was to evaluate the relative performance of these methods under the constraints of limited labeled data, rather than to introduce new modeling innovations. The experiments included the use of a large language model (GPT-4o) in a few-shot prompting setup and the fine-tuning of BERT-based models (MARBERT and SODA-BERT).

Few-Shot Prompting with GPT-4o:

The first approach utilized GPT-4o in a few-shot setting. Figure 1 shows a manually crafted prompt was designed for the sentiment classification task, where each sentiment class in each dialect was represented by two example tweets. This prompt was then used to classify a subset of test samples. While initial manual evaluation on selected examples indicated reasonable performance, the official test set evaluation produced an F1-score of

0.69, suggesting moderate effectiveness of this few-shot strategy without further domain adaptation.

dialect generalization and robustness. This result highlights the model's capacity to capture sentiment-relevant linguistic patterns that extend beyond the boundaries of a single dialect, despite the inherent phonological, lexical, and syntactic differences between Omani, Saudi, and Darija Arabic.



```
You are a sentiment classification expert. Given an Arabic text written in a specific dialect,
classify its sentiment into one of the following categories:
- "positive"
- "neutral"
- "negative"

Respond with **only** the category label.

Here are some examples:

Sentiment: neutral
Dialect: Darija
Text: فندق كايمشي حاله بالنسبة للثمن انا عن نفسي عادي أهم شي مكان نقي للنوم مادام أنا عازب أو متزوجين ما كنوصيش بهاد
الفندق يبقى وجه نظر والفندق مستوى نجوم و شكرا ليكم على الفرصة لي نعلق و نبدي رأي

Sentiment: neutral
Dialect: Darija
Text: فندق معقول بزاف بالنسبة للشقق فندق معقول بزاف قريب من المترو و المطار و مركز التسوق سيتي سنتر دير بس هو ممكن
يصنف فطور مش ربعة نجوم شوية زوين بس مش فخم بس

Sentiment: neutral
Dialect: Saudi
Text: حتى احسن قد كتبت عن العاج من قبل كان عندي في الفترة الاخيرة سبب لزيارتهم مره ثانية و تفاجأت يبدو ان المكان غير
الموظفين و جالسين يصلحون المكان حتى انه فيهافتتاح مطعم صغير قريبا وقد انعرض علي عرض من الدرجة الاولى على شقه
بغرفتين نوم هذا المكان سعره مناسب بالنسبة للتحسينات

Sentiment: neutral
Dialect: Saudi
Text: فندق قديم لكنه حلوه مره الفندق ما قد تجدد سكنت هنا في رحلة عمل لاني كنت مضطر لكنها كانت تجربة تونس مره

Sentiment: positive
Dialect: Saudi
Text: الاكل في كل الوجبات زين مرة مو ناقصه الا ثلاجة و مكعبات ثلج في كل دور و حبل غسيل داخلي مجهود ممتاز من كل
العاملين شكرا و اتمني ارجع مره ثانية

Sentiment: positive
Dialect: Saudi
Text: اطلق شي سافرنا كثير و ما قد لقينا مثل هذا القدر من الموده و المساعدة وي رجع الفضل الى الانسة تران و الطريقة الحلوة اللي
تعامل فيها طاقم الموظفين واح نرجع مستقبلا و قد علمنا كل اصدقاءنا عن كل شي حلو في هذا المكان

Sentiment: positive
Dialect: Darija
Text: وفيه واحد الفرقة دالتنشيط واعرة، كاين معاملة مزيانة ومحترمة لكاع الكليان

Sentiment: positive
Dialect: Darija
Text: فندق نقي بزاف و الاكل رائع، ولكن الاكل و الفندق نقي بزاف و التنظيم زوين بزاف، ولكن الفندق ماشي مناسب للعرب و
المصريين، الحمام ماشي فيه شطاف، كيدز كلوب ف ف فندق آخر و ماكايهضروش فيه العربية، ماشي ضرورة لولادك يكونوا هنا،
فالفندق فيه نسبة كبيرة من العمال الأجانب فالإستقبال و الكونسيرج، ماعرفتش واش ماكاينش مصريين ماكانش مز

Sentiment: negative
Dialect: Darija
Text: واخا كانت عندنا ريزيرفاسيون ما احترموهاش. وصلنا شوية دقائق قبل الوقت وما كان حتى عذر.

Sentiment: negative
Dialect: Darija
Text: الغرف ما منظمينش مزيان، وبعض الأدوات ما كايناش، بحال حذاء دورة المياه. يقدرو يديرو أحسن من هكا، خصوصاً مع الثمن
الغالي اللي كيطلبو دايماً.

Sentiment: negative
Dialect: Saudi
Text: العيب في هذا الفندق ان الغرف صغيرة و الاثاث قديم مره و لا فيه ثلاجة و لا غلاية

Sentiment: negative
Dialect: Saudi
Text: الغرف ماهي مرتبة زين و بعض الادوات ماهي موجودة مثل نعول دورة المياه يقدرون يكونون افضل من كذا مقابل السعر
العالي اللي يطلبونه

Now classify the following:

Dialect: {DIALECT}
Text: {TEXT}
Sentiment:

// where values between curly braces are variables
```

Figure 1: Prompt Structure used for Few-shot Experiment

Evaluation of SODA-BERT (Omani Fine-Tuned MARBERT):

The second approach involved evaluating SODA-BERT [1] , a custom model based on MARBERT that had been previously fine-tuned on sentiment data from the Omani dialect, it was directly applied to the Saudi and Darija tweets from the Ahasis test set without any further adaptation or fine-tuning. Surprisingly, SODA-BERT achieved an F1-score of 0.71, demonstrating notable cross-

Fine-Tuning MARBERT on Ahasis Training Set:

The final and most effective approach involved directly fine-tuning the original MARBERT model on the labeled training set provided for the task.

During preprocessing, a dialect identifier token was added at the beginning of each tweet (e.g., [DIALECT] text) to help the model distinguish between dialectal variations. The model was fine-

[1] SODA-BERT https://huggingface.co/mktr/SODA-BERT

tuned using the Hugging Face Transformers library with the following setup: a learning rate of 5e-5, cosine scheduler with warmup (1000 warmup steps), and AdamW optimizer configured with β_1=0.9 and β_2=0.98. Training was conducted for 8 epochs with a batch size of 16, gradient accumulation steps of 2, and label smoothing factor of 0.05 to enhance generalization. A macro F1-score was used as the primary evaluation metric, with early stopping based on validation performance. The maximum sequence length was set to 128 tokens. To improve training stability and efficiency, mixed precision (fp16) and gradient checkpointing were enabled, and a random seed of 42 was fixed for reproducibility. This fine-tuning procedure achieved the highest F1-score of 0.75 on the test set, confirming the value of domain-specific adaptation, especially in settings with multiple dialects and limited annotated data.

The comparative evaluation of these approaches highlights that while large LLMs like GPT-4o show promise in few-shot scenarios, transformer-based models fine-tuned on task-relevant data remain more effective for dialectal sentiment classification. Additionally, the cross-dialect performance of SODA-BERT offers valuable insights into the transferability of sentiment knowledge across closely related Arabic dialects.

4 Results and Discussion

This section presents and discusses the results of the sentiment classification system developed for the Ahasis shared task. The goal was to classify tweets written in Saudi and Darija dialects into positive, negative, or neutral categories. To address this task, a MARBERT model was fine-tuned on the official Ahasis training dataset and used to perform sentiment prediction on the provided test set. MARBERT, originally pre-trained on a large corpus of Arabic dialectal data, was selected for its strong performance on similar Arabic language tasks.

The official test set provided by the organizers contained 216 tweets, equally distributed between the two dialects. Table 1 summarizes the sentiment distribution within the test set:

Tabel 1: Sentiment Distribution in the Ahasis Shared Task Test Dataset

Dialect	Positive	Negative	Neutral
Darija	47	39	22
Saudi	42	37	29

And within the train set:

Tabel 2: Sentiment Distribution in the Ahasis Shared Task Train Dataset

Dialect	Positive	Negative	Neutral
Darija	154	168	108
Saudi	154	168	108

To address this task, the MARBERT model was fine-tuned on the provided training dataset. A preprocessing step was incorporated where a dialect-specific token was added at the beginning of each tweet to guide the model in differentiating between dialects. The experiments were conducted using the Hugging Face Transformers library, with hyperparameters adjusted through manual tuning based on validation performance.

During the training and validation phases, several experiments were carried out to optimize the model's performance. These experiments included varying learning rates, batch sizes, and epoch counts. The model demonstrated stable and consistent performance across different configurations, achieving its highest F1-score on the validation set with a learning rate of 2e-5, a batch size of 16, and 5 training epochs.

The final fine-tuned MARBERT model was then evaluated on the official Ahasis test set. It achieved an F1-score of 0.75, along with precision, recall, and accuracy values also equal to 0.75, indicating stable and consistent performance across key metrics. These results show a substantial improvement over the baseline system reported by Alharbi et al. (2025), which achieved an F1-score of 0.56 on the same task. Table 3 summarizes the final evaluation metrics.

Table 3. Performance Metrics of the Fine-Tuned MARBERT Model.

Metric	Value
F1-score	0.75
Accuracy	0.75
Precision	0.75
Recall	0.75

Balanced Accuracy	0.746

For context, Table 4 lists the top three performing teams in the Ahasis shared task based on their F1 scores.

Table 4. Top 3 Team Rankings in the Ahasis Shared Task

Rank	**Team**	**Score - F1**
1	**Hend**	**0.81**
2	ISHFMG_TUN	0.79
3	LBY	0.79

5. Conclusion

Following the conclusion of the Ahasis shared task, our proposed system secured 8th place among all participating teams. This result highlights the model's robustness and its ability to handle the complexities of under-resourced dialects such as Saudi and Darija. The system's strong performance can be attributed to the effective application of task-specific fine-tuning and the inclusion of dialect-aware preprocessing, which helped the model differentiate linguistic patterns Among dialects. These findings reaffirm that carefully adapted transformer-based models remain a dependable baseline for Arabic dialect sentiment analysis, especially in low-resource scenarios. While our approach did not introduce novel modeling techniques, it offers practical insights into the capabilities and limitations of standard methods when applied thoughtfully to challenging dialectal data.

References

Ahasis Shared Task Organizers. 2025. "The Ahasis Shared Task on Saudi and Darija Sentiment." *Workshop on Arabic Shared Tasks*, pp. 12–18.

Ahmed Abdel-Salam and Hamdy Mubarak. 2023. "Arabic dialect identification and sentiment analysis using prompt-tuning and MARBERT." In *Proceedings of the Arabic NLP 2023 Shared Task*, Association for Computational Linguistics, Doha, Qatar, pp. 24–32.

Mohamed A. Abdelwahab, Ahmed Ramy, and Hala Abou-Assaleh. 2022. *Explainable sentiment classification in Arabic tweets using LIME*. Procedia Computer Science, vol. 202, pp. 1223–1232.

Al-Twairesh, N., Al-Khalifa, H. S., & Al-Salman, A. (2018). *Sentiment Analysis of Arabic Tweets: Feature Engineering and a Hybrid Approach. arXiv preprint arXiv:1805.08533.*Atoum, J. O., & Nouman, M. (2019). *Sentiment Analysis of Arabic Jordanian Dialect Tweets. International Journal of Advanced Computer Science and Applications (IJACSA)*, 10(2), 107–112. https://doi.org/10.14569/IJACSA.2019.0100234

Mohammad Alsarhan and Houda Bouamor. 2023. "Sentence embeddings for Gulf dialect sentiment analysis." In *Proceedings of the Arabic NLP 2023 Shared Task*, Association for Computational Linguistics, Doha, Qatar, pp. 102–109.

Ali A. Alsayat and Ahmed E. El-Sayed. 2022. *Sentiment analysis of Saudi dialect tweets using machine learning techniques*. IEEE Access, vol. 10, pp. 74830–74838.

Antoun, W., Baly, F., & Hajj, H. (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)*, pp. 9–15. Available: https://aclanthology.org/2020.osact-1.2/

Nizar Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool.

Hamdy Mubarak, Youssef Al-Khamissi, and Kareem Darwish. 2023. "AraGPT: An Arabic generative pre-trained transformer for NLP." In *Proceedings of ACL 2023*, Association for Computational Linguistics, Toronto, Canada, pp. 240–249. http://aclweb.org/anthology/2023.acl-main.24

Bo Pang and Lillian Lee. 2008. "Opinion mining and sentiment analysis." *Foundations and Trends in Information Retrieval*, vol. 2(1–2), pp. 1–135.

Ahmed Zahran and Yasmine Elglaly. 2024. "Building dialectal Arabic sentiment resources: Saudi and Moroccan challenges." In *Proceedings of the 2nd Arab Shared Task Workshop*, pp. 56–64.