AHaSIS-ST 2025

Proceedings
of the Shared Task on Sentiment Analysis on Arabic Dialects in the Hospitality Domain: A Multi-Dialect Benchmark

associated with

The 15th International Conference on Recent Advances in Natural Language Processing RANLP'2025

Edited by Maram I. Alharbi, Salmane Chafik, Saad Ezzini, Ruslan Mitkov
Tharindu Ranasinghe and Hansi Hettiarachchi

12 September, 2025
Varna, Bulgaria

Shared Task on Sentiment Analysis on Arabic Dialects
in the Hospitality Domain: A Multi-Dialect Benchmark
Associated with the International Conference
Recent Advances in Natural Language Processing
RANLP'2025

PROCEEDINGS

Varna, Bulgaria
12 September 2025

Online ISBN 978-954-452-109-7

Designed by INCOMA Ltd.
Shoumen, BULGARIA

Preface

Welcome to the Sentiment Analysis on Arabic Dialects in the Hospitality Domain (AHaSIS) Shared Task, held on September 12, 2025, in conjunction with the 15th International Conference on Recent Advances in NLP (RANLP 2025) in Varna, Bulgaria.

With the rapid digital transformation of the hospitality industry across the Arab world, user-generated content such as hotel reviews has become a critical source of insight for service enhancement. However, the complexity of Arabic and its many dialects poses unique challenges for Natural Language Processing (NLP), particularly in sentiment analysis. The AHaSIS shared task addresses this need by focusing on sentiment detection across multiple Arabic dialects, specifically within the hospitality domain.

This shared task introduces a novel, sentiment-balanced dataset consisting of over 1000 hotel reviews, originally authored in Modern Standard Arabic (MSA) and professionally translated into two major dialects: Saudi Arabic and Moroccan Darija. The translations were validated by native speakers to ensure both linguistic and sentiment fidelity, making this dataset a valuable resource for developing and benchmarking dialect-aware sentiment analysis systems.

The AHaSIS task attracted widespread interest from the NLP community, with over 40 teams registering to participate. Of these, 12 teams successfully submitted systems during the evaluation phase. The top-performing system achieved an F1 score of 0.81, demonstrating both the promise and the persistent challenges of robust sentiment analysis in Arabic dialects.

We believe the outcomes of the AHaSIS shared task contribute significantly to the advancement of Arabic NLP, especially for underrepresented dialectal varieties in practical, real-world domains such as hospitality. The shared task encourages further research into multi-dialect processing and highlights the importance of dialect-specific resources for effective sentiment analysis.

We would like to express our sincere gratitude to all participating teams, reviewers, and the organizing committee for their contributions and dedication. We hope that these proceedings will support and inspire future work on sentiment analysis and dialectal Arabic NLP.

Maram Alharbi, Shared Task Chair, on behalf of the AHaSIS organizing team.

Organizing Committee

Maram Alharbi, Lancaster University, UK
Salmane Chafik, Mohammed VI Polytechnic University, Morocco
Saad Ezzini, King Fahd University of Petroleum and Minerals, Saudi Arabia
Ruslan Mitkov, Lancaster University, UK
Tharindu Ranasinghe, Lancaster University, UK
Hansi Hettiarachchi, Lancaster University, UK

Table of Contents

Shared Task Program

Friday, September 12, 2025

9:00–9:05 ***Welcome and Opening Remarks***

9:05–9:15 *AHaSIS: Shared Task on Sentiment Analysis for Arabic Dialects*
Maram I. Alharbi, Salmane Chafik, Saad Ezzini, Ruslan Mitkov, Tharindu Ranasinghe and Hansi Hettiarachchi

Session 1

9:15–9:30 *iWAN-NLP at AHaSIS 2025: A Stacked Ensemble of Arabic Transformers for Sentiment Analysis on Arabic Dialects in the Hospitality Domain*
Hend Al-Khalifa

9:30–9:45 *Fine-tuning AraBert model for arabic sentiment detection*
Mustapha Jaballah, Dhaou Ghoul and Ammar Mars

9:45–10:00 *Enhancing Arabic Dialectal Sentiment Analysis through Advanced Data Augmentation Techniques*
Md. Rafiul Biswas and Wajdi Zaghouani

10:00–10:15 *Ahasis Shared Task: Hybrid Lexicon-Augmented AraBERT Model for Sentiment Detection in Arabic Dialects*
Shimaa Amer Ibrahim, Mabrouka Bessghaier and Wajdi Zaghouani

10:15–10:30 *Lab17 @ Ahasis Shared Task 2025: Fine-Tuning and Prompting techniques for Sentiment Analysis of Saudi and Darija Dialects*
Al Mukhtar Al Hadhrami, Firas Al Mahrouqi, Mohammed Al Shaaili and Hala Mulki

10:30–10:45 *Dialect-Aware Sentiment Analysis for Ahasis Challenge*
Hasna Chouikhi and Manel Aloui

10:45–11:00 ***Coffee Break***

Friday, September 12, 2025 (continued)

Session 2

AHaSIS: Shared Task on Sentiment Analysis for Arabic Dialects

Maram Alharbi[1,2], **Salmane Chafik**[3], **Saad Ezzini**[4], **Ruslan Mitkov**[1]
Tharindu Ranasinghe[1] **and Hansi Hettiarachchi**[1]
[1]School of Computing and Communications, Lancaster University, UK
[2]Jazan University, Saudi Arabia
[3]Mohammed VI Polytechnic University, Morocco
[4]King Fahd University of Petroleum and Minerals, Saudi Arabia
m.i.alharbi@lancaster.ac.uk

Abstract

The hospitality industry in the Arab world increasingly relies on customer feedback to shape services, driving the need for advanced Arabic sentiment analysis tools. To address this challenge, the Sentiment Analysis on Arabic Dialects in the Hospitality Domain shared task focuses on Sentiment Detection in Arabic Dialects. This task leverages a multi-dialect, manually curated dataset derived from hotel reviews originally written in Modern Standard Arabic (MSA) and translated into Saudi and Moroccan (Darija) dialects. The dataset consists of 538 sentiment-balanced reviews spanning positive, neutral, and negative categories. Translations were validated by native speakers to ensure dialectal accuracy and sentiment preservation. This resource supports the development of dialect-aware NLP systems for real-world applications in customer experience analysis. More than 40 teams have registered for the shared task, with 12 submitting systems during the evaluation phase. The top-performing system achieved an F1 score of 0.81, demonstrating the feasibility and ongoing challenges of sentiment analysis across Arabic dialects.

1 Introduction

Arabic Sentiment Analysis (ASA) has become an increasingly prominent field within Natural Language Processing (NLP), spurred by the growing volume of Arabic content across digital platforms and the pressing need for automated systems to gauge public opinion. In contrast to high-resource languages, ASA continues to face enduring challenges due to the linguistic complexity of Arabic, its diglossic nature, and the considerable variation across regional dialects (Habash et al., 2013). These challenges are particularly evident in informal domains such as social media and hospitality, where sentiment expressions differ significantly across dialects.

To date, the majority of available resources for ASA have concentrated on Modern Standard Arabic (MSA), offering limited applicability to dialectal variants (Aladeemy et al., 2024). Consequently, models trained on MSA frequently struggle to generalise across dialects, leading to diminished performance in practical settings (Khrisat and Al-Harthy, 2015). Additionally, the development of robust, dialect-sensitive models has been hindered by a notable lack of high-quality, annotated datasets.

In response to these limitations, we present the Ahasis 2025 Shared Task, which seeks to advance sentiment classification techniques across Arabic dialects within the hospitality domain. This shared task provides a balanced dataset comprising hotel reviews written in Saudi Arabic and Moroccan Darija, each annotated with sentiment labels. Participants are invited to explore both traditional and neural classification approaches under conditions of limited training data. The task aims to evaluate the effectiveness of various modelling strategies in identifying sentiment from dialect-rich, user-generated content.

The remainder of this paper is organised as follows: Section 2 reviews the relevant literature; Section 3 details the shared task and its setup; Section 4 describes the dataset; Section 5 presents the evaluation results; and finally, the paper concludes with key findings and outlines future directions.

2 Related Work

Arabic sentiment analysis has witnessed growing attention in recent years, with early studies laying the foundation by addressing the lack of dialect-specific annotations and lexical resources (Nabil et al., 2015). Aladeemy et al. (2024) critically reviewed the state of sentiment annotation in Arabic dialects, highlighting the prevalence of manual labelling techniques and the limited use of auto-

Proceedings of the Shared Task on Sentiment Analysis for Arabic Dialects (AHaSIS) associated with RANLP 2025, pages 1–6, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-109-7-001

mated methods due to a shortage of robust linguistic resources. Their findings emphasise that machine learning approaches dominate the field, while lexicon-based systems remain underutilised.

Recent literature has placed emphasis on tackling dialectal diversity, recognising that Arabic dialects differ significantly in syntax, morphology, and vocabulary. A systematic review by Matrane et al. (2023) identified key preprocessing stages, such as normalisation, feature extraction, and sentiment tagging, as decisive factors in improving classification performance. The review also underscored the importance of handling negation and morphological variation, both of which are vital to interpreting sentiment in dialectal contexts.

Deep learning architectures, including convolutional neural networks (CNNs) and recurrent models like LSTM (Hochreiter and Schmidhuber, 1997) and GRU (Chung et al., 2014), have shown strong results in Arabic sentiment tasks (Baali and Ghneim, 2019). However, preprocessing remains a critical bottleneck. Guellil et al. (2020) stressed the necessity of standardised pipelines to improve performance consistency across tasks and datasets.

In parallel, researchers have explored cross-lingual methods to augment Arabic sentiment resources. Saadany and Orasan (2020) investigated the preservation of sentiment polarity in neural machine-translated Arabic reviews and identified frequent distortions introduced by automated translation tools. Similarly, Poncelas et al. (2020) examined the impact of machine translation on downstream sentiment classification, revealing that models trained on original data outperform those trained on translated corpora, especially in sentiment-sensitive applications.

Finally, while most progress has been made in MSA, Aladeemy et al. (2024) emphasise that Arabic dialects remain underrepresented in sentiment analysis research. They call for a shift towards developing dialect-aware resources and models that address the linguistic variation inherent to Arabic. The Ahasis shared task responds to this call by offering a domain-specific, multi-dialectal dataset and encouraging participants to experiment with resource-efficient and generative learning paradigms.

3 Task Description

3.1 Sentiment Detection in Arabic Dialects

The Ahasis 2025 Shared Task centres on sentiment analysis within the hospitality domain, specifically targeting hotel reviews written in regional Arabic dialects. Given Arabic's linguistic richness, marked by the coexistence of MSA and a wide range of spoken dialects, sentiment classification presents notable challenges. Dialects vary considerably in morphology, syntax, and vocabulary, and this variability is further amplified in informal user-generated content, where sentiment is often conveyed through idiomatic or region-specific expressions.

Participants are required to classify hotel reviews into one of three sentiment categories: *positive*, *neutral*, or *negative*. The data comprises user reviews in Saudi and Darija dialects. The task evaluates participants' ability to build models that can generalise across dialects while maintaining high accuracy in nuanced sentiment interpretation. This shared task explicitly encourages the development of techniques that are resilient to linguistic variation and reflective of real-world text usage in the hospitality sector.

3.2 Resources and Evaluation

Participants will be provided with a bi-dialect annotated dataset of hotel reviews. The task permits the use of external resources, including pre-trained encoders, large language models, and data augmentation techniques, allowing for a wide exploration of modelling strategies.

The primary evaluation metric is the **F1-score**, computed over the three sentiment classes. In addition, secondary analyses will include:

- **Dialect-Specific Performance:** Evaluating performance across Saudi and Darija dialects separately.
- **Error Categorisation:** Analysing model errors in terms of sentiment misclassification, dialectal confusion, or ambiguous content.

4 Data

The shared task provides a bi-dialect Arabic sentiment dataset specifically designed for the hospitality domain. The dataset comprises hotel review sentences in two Arabic dialects: Saudi and Moroccan Darija. Each review is annotated with a sentiment label (positive, neutral, or negative), enabling both dialect-specific and cross-dialect sentiment analysis.

The original data was derived from the ABSA-Hotels dataset released as part of the Arabic track of SemEval-2016 (Pontiki et al., 2016; Al-Dabet et al., 2021). This dataset consists of Arabic hotel reviews sourced from platforms such as Booking.com and TripAdvisor. The base data, originally in MSA, was extensively preprocessed and refined following the approach described in (Alharbi et al., 2025).

4.1 Dataset Structure

The dataset released for this shared task is organized into training and test splits, both covering two Arabic dialects: Saudi and Moroccan Darija. Each instance in the dataset represents a hotel review sentence. The training set includes sentiment annotations, while the test set is used for evaluation and does not expose the sentiment labels.

Split	Entries	Dialects	Sentiment Labels
Train	860	Saudi, Darija	Positive, Neutral, Negative
Test	216	Saudi, Darija	N/A (to be predicted)

Table 1: Structure and statistics of the shared task dataset.

Participants are required to use the provided fields in the test set to predict sentiment labels, ensuring their models generalise well across dialects. The task emphasises robustness to dialectal variation and sentiment nuance, with all reviews grounded in real-world user feedback from the hospitality sector.

5 Results and Analysis

The Ahasis shared task attracted a diverse set of participants, showcasing a range of modelling techniques and domain-specific innovations. Participants engaged with the challenge of accurately classifying sentiment in two dialects, Saudi and Darija, using both fine-tuned transformer models and large language models (LLMs) with prompt engineering strategies.

With sentiment classification task, the competition provided a realistic, low-resource benchmark reflective of linguistic variability and cultural nuance in Arabic dialects. The dataset's domain focus on hospitality reviews added complexity through indirect sentiment cues, politeness strategies, and dialect-specific idioms. Teams employed diverse strategies, including fine-tuning pretrained models such as MARBERTv2 (Abdul-Mageed et al., 2021), DarijaBERT (Gaanoun et al., 2024), and AraBERT (Abdul-Mageed et al., 2021) variants, or leveraging zero-shot capabilities of LLMs like Gemini Pro.

5.1 Participating Teams and Final Rankings

A total of 12 teams submitted systems for the test phase. Table 2 presents the final leaderboard based on micro-averaged F1 scores.

The top-performing system, submitted by Team Hend, achieved an F1 score of 0.81, followed closely by ISHFMG_TUN and LBY with scores of 0.79.

The results show a tight clustering of top scores between 0.73 and 0.81, with strong performances across a variety of modelling strategies, including fine-tuned transformer models, few-shot LLM prompting, and hybrid lexical-embedding methods.

5.2 Team Description

1. Hend (iWAN-NLP): The iWAN-NLP team participated in the AHaSIS 2025 shared task with a transformer-based ensemble system designed for sentiment analysis across Arabic dialects. Their approach combined three pre-trained models, MARBERTv2 (Abdul-Mageed et al., 2021), SaudiBERT (Qarah, 2024), and DarijaBERT (Gaanoun et al., 2024), each fine-tuned using stratified 5-fold cross-validation. The ensemble was built by averaging logits across folds and models, leveraging model diversity to improve robustness. Training enhancements included label smoothing, mixed-precision training, early stopping, and learning rate warmup. This system achieved a micro F1 score of 0.81, ranking first among all participants.

2. ISHFMG_TUN: This team tackled the sentiment analysis task by fine-tuning the AraBERTv02 model (Abdul-Mageed et al., 2021), a pre-trained Arabic language model optimised for social media text. Their approach incorporated several fine-tuning strategies, including freezing lower transformer layers, applying class weighting to address imbalance, and tuning dropout and learning rate schedules. They trained the model using validated on both Saudi and Darija dialects. Without relying on external data, their system achieved a micro F1 score of 0.7916, ranking second in the AHaSIS 2025 shared task.

3. LBY: The LBY team tackled the AHaSIS 2025 shared task by fine-tuning six pre-trained Arabic transformer models, including bert-base-arabert,

Submission ID	Codalab Username	Team Name	Test Phase Micro-F1	Rank
281611	hend_suliman	Hend (iWAN-NLP)	0.81	1
282197	ishfmgtun	ISHFMG_TUN	0.79	2
282404	nwesri	LBY	0.79	3
282005	hasnachouikhi	LahjaVision	0.77	4
282408	msmadi	AraNLP	0.76	5
282490	ahmedabdou	MucAI	0.76	6
280604	shimaa	MARSAD	0.75	7
281739	almktr	Lab17	0.75	8
281362	salwas	BirLee	0.75	9
282386	mabrouka4	MARSAD AI	0.74	10
282374	mlubbad	Lubbad	0.74	11
282445	zarnoufi	MAPROC	0.73	12

Table 2: Ahasis Shared Task Test Phase results ranked by Micro-F1.

bert-base-arabertv02-twitter, bert-large-arabertv02-twitter, MARBERTv2, bert-base-qarib, and DarijaBERT. Their focus was on assessing model performance across both Saudi and Darija dialects. Through a series of dialect-specific and combined-dialect experiments, MARBERTv2 emerged as the top-performing model in their setup. The team emphasised robust training strategies and careful hyperparameter tuning over ensembling, achieving an F1 score of 0.79, and securing third place in the official evaluation.

4. LahjaVision: Representing a dialect-focused approach to Arabic sentiment analysis, the LahjaVision team developed a dialect-aware system that leveraged the QARiB transformer model, enriched with specialised dialect embeddings and custom preprocessing for Saudi and Darija Arabic. Their methodology incorporated discriminative fine-tuning, focal loss, and dialect-specific normalisation to better capture sentiment expressions. By embedding dialect information into the model architecture, they achieved notable improvements over baseline and non-dialect-aware systems. Their final system attained a micro F1 score of 0.77, securing fourth place in the AHaSIS 2025 shared task.

5. AraNLP: Competing in the Ahasis 2025 shared task on sentiment analysis for Arabic hotel reviews, the AraNLP team proposed a hybrid deep learning architecture combining the transformer-based AraELECTRA (Antoun et al., 2021) model with classical TF-IDF features. This design aimed to capture both contextual semantics and important lexical cues, particularly for dialects like Saudi and Darija. The model was trained with minimal preprocessing to preserve dialectal expressions and used a feature fusion mechanism to integrate embeddings and lexical vectors. AraNLP achieved a micro F1-score of 76%, securing fifth place among all participants.

6. MucAI: The MucAI team approached the AHaSIS 2025 shared task using an innovative few-shot prompting strategy with GPT-4o for Arabic sentiment analysis. They explored zero-shot, static, and adaptive prompting methods, with their final system dynamically retrieving the most semantically similar examples via kNN search over AraBERT embeddings. Each selected example was paired with a chain-of-thought explanation, forming a tailored prompt per review. This adaptive prompting approach significantly improved performance, especially for neutral sentiment cases. The system achieved a micro F1-score of 76%, outperforming static and zero-shot setups and earning sixth place in the shared task.

7. MARSAD: This team tackled Arabic sentiment analysis in the hospitality domain by applying structured data augmentation to enhance performance in low-resource dialectal settings. They combined three techniques, paraphrasing via FANAR API, pattern-based sentence generation, and domain-specific word substitution, while retaining dialect-specific linguistic cues. Their approach utilized AraBERT-Large-v02 fine-tuned on both original and augmented data. The resulting system achieved a micro F1-score of 0.75, securing 8th place in the shared task.

8. LAB17: This team combined generative and transformer-based strategies for sentiment analysis in Arabic dialect hotel reviews. They applied few-shot prompting with GPT-4o and fine-tuned transformer models, including MARBERT and its Omani-dialect variant, SODA-BERT. While GPT-4o reached a micro F1-score of 0.69, their fine-tuned MARBERT model outperformed all with a micro F1-score of 0.75, securing 8th place in the shared task.

9. BirLee: The BirLee team focused on sentiment analysis for Saudi and Darija dialects by fine-tuning CAMeLBERT-DA with both hotel reviews and a newly curated Saudi proverbs dataset. Their model achieved a micro F1-score of 0.75, outperforming Arabic-centric large language models like Allam 0.70, ACeGPT 0.68, and Jais 0.65 in zero-shot settings. Their results emphasise the effectiveness of domain-specific fine-tuning over zero-shot strategies in dialectal Arabic sentiment analysis.

10. MARSAD AI: This team tackled sentiment analysis in Arabic dialect hotel reviews through a hybrid model approach. Their system combined contextual embeddings from AraBERT with a custom-built sentiment lexicon tailored to Saudi and Darija dialects. To overcome data scarcity, they implemented two augmentation strategies: probabilistic lexical perturbation and paraphrasing using AraT5. This enriched and diversified the training data. The resulting hybrid model significantly outperformed the baseline AraBERT-only setup, achieving an F1 score of 0.74

11. Lubbad: Lubbad tackled the sentiment analysis task using the Gemini Pro 1.5 large language model. Instead of retraining, the team employed dialect-specific prompt engineering with real-time batch inference. The approach incorporated sarcasm detection, dialect labelling, and custom zero/few-shot prompts optimized for Saudi and Darija dialects. The system achieved a micro F1-score of 0.7361, ranking 10th in the Ahasis Shared Task.

12. MAPROC: This team participated in the Ahasis shared task using the SetFit framework, a few-shot learning technique based on fine-tuning sentence transformers. They employed the Arabic-SBERT-100K model and experimented with limited examples per class, ultimately using 64 examples per sentiment category for contrastive fine-tuning. Their approach demonstrated the potential of data-efficient modeling in low-resource dialectal sentiment classification and achieved a micro F1-score of 0.73, placing 12th on the leaderboard.

6 Conclusion and Future Work

In conclusion, the Ahasis 2025 Shared Task marks a significant step forward in advancing Arabic sentiment analysis, particularly in addressing the challenges posed by dialectal variation in the hospitality domain. By focusing on sentiment detection in Saudi and Darija dialects, the task has created a valuable benchmark for evaluating NLP systems under low-resource, real-world conditions. The participation of diverse teams employing a range of methodologies, from transformer fine-tuning to few-shot prompting, has yielded meaningful insights into effective modelling strategies for dialectal sentiment classification.

The results highlight the impressive performance of several teams, most notably iWAN-NLP team, whose ensemble of fine-tuned BERT-based models achieved the highest F1 score. This underscores the critical role of both model sophistication and dialect-specific data curation in achieving high performance. Moreover, systems that integrated domain knowledge, robust preprocessing, or adaptive prompting techniques also demonstrated strong capabilities, reflecting the importance of combining linguistic insight with technical innovation.

Looking ahead, future work in Arabic dialect sentiment analysis could explore broader dialectal coverage and task extensions such as aspect-based sentiment analysis or emotion detection. Continued development of pre-trained dialectal models and domain-specific embeddings will also be essential for improving generalisability and robustness.

The Ahasis Shared Task has laid a foundation for future research in Arabic dialect NLP, promoting collaboration and innovation in a field that remains under-represented yet highly impactful. By advancing the development of inclusive, dialect-aware NLP systems, this shared task contributes to broader efforts in enhancing the linguistic diversity, cultural relevance, and real-world applicability of sentiment analysis technologies in the Arab world.

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

Saja Al-Dabet, Sara Tedmori, and Mohammad AL-Smadi. 2021. Enhancing arabic aspect-based sentiment analysis using deep learning models. *Computer Speech & Language*, 69:101224.

Amani A. Aladeemy, Ali Alzahrani, Mohammad H. Algarni, Saleh Nagi Alsubari, Theyazn H.H. Aldhyani, Sachin N. Deshmukh, Osamah Ibrahim Khalaf, Wing-Keung Wong, and Sameer Aqburi. 2024. Advancements and challenges in arabic sentiment analysis: A decade of methodologies, applications, and resource development. *Heliyon*, 10(21):e39786.

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

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. AraELECTRA: Pre-training text discriminators for Arabic language understanding. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 191–195, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

M. Baali and N. Ghneim. 2019. Emotion analysis of arabic tweets using deep learning approach. *Journal of Big Data*, 6(1):89.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Kamel Gaanoun, Abdou Mohamed Naira, Anass Allak, and Imade Benelallam. 2024. Darijabert: a step forward in nlp for the written moroccan dialect. *International Journal of Data Science and Analytics*, pages 1–13.

I. Guellil, A. Adeel, F. Azouaou, F. Benali, A. Hachani, and A. Hussain. 2020. Sentiment analysis of arabic dialects. Technical report, Ecole Superieure des Sciences Appliquées d'Alger, Laboratoire des Méthodes de Conception des Systèmes, Ecole nationale Supérieure d'Informatique, & University of Stirling.

Nizar Habash, Owen Rambow, and Ryan Roth. 2013. Morphological analysis and generation for arabic dialects: Challenges and solutions. *Computational Linguistics*, 39(3):453–495.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

A. Khrisat and Z. Al-Harthy. 2015. Arabic dialects and classical arabic language. *Advances in Social Sciences Research Journal*, 2(3):254–260.

Y. Matrane, F. Benabbou, and N. Sael. 2023. A systematic literature review of arabic dialect sentiment analysis. *Journal of King Saud University-Computer and Information Sciences*, 35(6):101570.

M. Nabil, M. Aly, and A. F. Atiya. 2015. Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 697–703.

Alberto Poncelas, Pintu Lohar, James Hadley, and Andy Way. 2020. The impact of indirect machine translation on sentiment classification. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 78–88, Virtual. Association for Machine Translation in the Americas.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Faisal Qarah. 2024. Saudibert: A large language model pretrained on saudi dialect corpora. *arXiv preprint arXiv:2405.06239*.

Hadeel Saadany and Constantin Orasan. 2020. Is it great or terrible? preserving sentiment in neural machine translation of Arabic reviews. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 24–37, Barcelona, Spain (Online). Association for Computational Linguistics.

iWAN-NLP at AHaSIS 2025: A Stacked Ensemble of Arabic Transformers for Sentiment Analysis on Arabic Dialects in the Hospitality Domain

Hend S. Al-Khalifa
iWAN Research Group, College of Computer and Information Sciences
King Saud University, Riyadh, Saudi Arabia
hendk@ksu.edu.sa

This paper details the iWAN-NLP system developed for participation in the AHaSIS 2025 shared task, "Sentiment Analysis on Arabic Dialects in the Hospitality Domain: A Multi-Dialect Benchmark." Our approach leverages a multi-model ensemble strategy, combining the strengths of MARBERTv2, SaudiBERT, and DarijaBERT. These pre-trained Arabic language models were fine-tuned for sentiment classification using a 5-fold stratified cross-validation methodology. The final predictions on the test set were derived by averaging the logits produced by each model across all folds and then averaging these combined logits across the three models. This system achieved a macro F1-score of 0.8055 (~0.81) on the official evaluation dataset and a cross-validated macro F1-score of 0.8513 (accuracy 0.8628) on the training set. Our findings highlight the effectiveness of ensembling regionally adapted models and robust cross-validation for Arabic sentiment analysis in the hospitality domain, ultimately securing first place in the AHaSIS 2025 shared task.

1 Introduction

The proliferation of user-generated content, particularly in the hospitality sector through reviews, provides a rich source of opinions and sentiments. For the Arabic language, this content is characterized by a complex interplay of Modern Standard Arabic (MSA) and numerous regional dialects. These dialects often exhibit substantial lexical, syntactic, and morphological divergence, posing considerable hurdles for Natural Language Processing (NLP) systems. The AHaSIS 2025 shared task, "Sentiment Analysis on Arabic Dialects in the Hospitality Domain: A Multi-Dialect Benchmark," aims to foster research in this area by providing a benchmark for evaluating systems on Arabic sentiment analysis within the context of hotel reviews, which can span MSA and various dialects.

Effective sentiment analysis of Arabic hotel reviews offers significant value for businesses in understanding customer satisfaction and for travelers in making informed decisions. However, the linguistic diversity, coupled with the nuances of sentiment expression (e.g., sarcasm, implicit feedback), makes this a challenging task. Transformer-based language models such as BERT (Devlin et al., 2019), have shown remarkable success in NLP by learning rich contextual representations, leading to the development of several Arabic-specific models.

In this paper, we present the iWAN-NLP system. Our system architecture is built upon an ensemble of multiple transformer models: MARBERTv2, chosen for its training on diverse Arabic social media; SaudiBERT, for its specialization in the Saudi dialect; and DarijaBERT, for its focus on North African Darija. Each base model was fine-tuned using 5-fold stratified cross-validation. Key training enhancements included a custom Hugging Face Trainer, label smoothing (0.1) via nn.CrossEntropyLoss, early stopping, FP16 mixed-precision training, and learning rate warmup. The final system predictions were generated by averaging model logits across folds and then across models. This approach yielded a macro F1-score of 81.0% on the

Proceedings of the Shared Task on Sentiment Analysis for Arabic Dialects (AHaSIS) associated with RANLP 2025, pages 7–13, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-109-7-002

official test set and a cross-validated macro F1-score of 85% on the training set. Notably, this performance earned our iWAN-NLP system the first-place ranking in the AHaSIS 2025 (Alharbi, Ezzini, et al., 2025) shared task on Sentiment Analysis on Arabic Dialects in the Hospitality Domain.

The remainder of this paper is organized as follows: Section 2 discusses related work. Section 3 describes the dataset. Section 4 details our methodology, including preprocessing, models, training, and ensemble strategy. Section 5 presents results. Section 6 offers an error analysis. Section 7 concludes and suggests future work.

2 Related Work

Sentiment analysis in Arabic has garnered significant attention in the NLP community. Early efforts often relied on lexicon-based approaches (Abdul-Mageed et al., 2012) or traditional machine learning algorithms like Support Vector Machines (SVMs) and Naive Bayes, typically using n-gram features or bag-of-words representations (El-Halees, 2011). While these methods provided initial breakthroughs, their performance was often limited by the morphological richness of Arabic and the challenges posed by dialectal variations.

The advent of deep learning, particularly Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), brought about improvements in capturing sequential and local features in text. However, the introduction of transformer-based models (Vaswani et al., 2017), especially BERT (Devlin et al., 2019), revolutionized the field. These models, pre-trained on massive text corpora, learn powerful contextual embeddings that have proven highly effective for a wide array of downstream tasks.

Several BERT-based models have been specifically developed for the Arabic language. AraBERT (Antoun et al., 2020) was one of the pioneering efforts, demonstrating strong performance on various Arabic NLP benchmarks. MARBERT (Abdul-Mageed et al., 2021), and its successor MARBERTv2, were trained on a significantly larger and more diverse dataset, including a substantial amount of Arabic social media text containing dialectal content, making them particularly well-suited for tasks like the AHaSIS challenge. Recognizing the limitations of general Arabic models when faced with specific dialects, researchers have also developed models tailored to regional variations. For instance, SaudiBERT (Qarah, 2024) was pre-trained exclusively on a large corpus of Saudi dialectal text, aiming to capture its unique linguistic characteristics. Similarly, DarijaBERT (Gaanoun et al., 2024) represents a significant step in NLP for the Moroccan Arabic dialect "Darija", providing the first set of BERT models specifically for this dialect.

Ensemble methods have a long history of improving machine learning model performance by combining the predictions of multiple individual learners. In NLP, ensembling has been successfully applied to various tasks, including sentiment analysis (Salah et al., 2019). While various ensemble techniques exist, including simple averaging/voting and more complex methods like stacking (Wolpert, 1992), the core idea is to leverage the diversity of multiple models. (Al Shamsi & Abdallah, 2023) explored ensemble techniques in the context of Arabic sentiment analysis, highlighting their potential for enhanced robustness.

To further optimize the training of large transformer models, several techniques have become standard practice. Label smoothing (Müller et al., 2019) is a regularization technique that prevents the model from becoming too confident in its predictions, which can improve generalization. Early stopping is widely used to prevent overfitting by monitoring performance on a validation set and halting training when performance ceases to improve. FP16 mixed-precision training (Micikevicius et al., 2018) allows for faster training and reduced memory footprint by using 16-bit floating-point numbers for certain operations, often without sacrificing performance. Learning rate warmup (Goyal et al., 2018), where the learning rate is gradually increased at the beginning of training, helps stabilize the training process, especially for large models and batch sizes.

Our work builds upon these advancements by combining multiple state-of-the-art Arabic transformer models, including those specialized for dialects, within an averaging ensemble framework, and by employing a suite of modern training enhancements to maximize performance.

3 Data

The AHaSIS 2025 shared task organizers provided a dataset consisting of Arabic hotel reviews. Each review was annotated with one of three sentiment labels: *positive, negative, or neutral*. The dataset was pre-split into official training and test sets. For our experiments, we strictly adhered to these provided splits and did not incorporate any external datasets or perform external data augmentation.

Labels (positive, negative, neutral) were mapped to numeric IDs for model processing. Tokenization was performed using each base model's associated pre-trained tokenizer. A key aspect of our data handling was the use of 5-fold Stratified K-Fold Cross-Validation on the training set. This approach ensures that each fold maintains a similar class distribution to the overall training set, providing a more reliable estimate of model performance during development and robust out-of-sample predictions from each fold for evaluating ensemble strategies.

4 Methodology

Our system for Arabic sentiment classification employs a multi-model ensemble approach, with each base model fine-tuned using a rigorous cross-validation strategy and enhanced training techniques.

4.1 Preprocessing

The preprocessing steps applied to the review data were minimal:

1. **Label Mapping:** The textual sentiment labels ('positive', 'negative', 'neutral') were mapped to numerical IDs (0, 1, 2).

2. **Tokenization:** Each review text was tokenized using the specific pre-trained tokenizer associated with the respective base model (MARBERTv2, SaudiBERT, DarijaBERT). This ensures that the input format matches what each model expects from its pre-training phase.

No further extensive preprocessing steps such as stopword removal, emoji normalization, or detailed punctuation cleaning were performed, relying on the inherent capabilities of the transformer models to process relatively raw text.

4.2 Base Models

We utilized three pre-trained Arabic transformer models from the HuggingFace Transformers library (Wolf et al., 2020) as our base learners:

- **MARBERTv2 (UBC-NLP/MARBERTv2)[1]:** Chosen for its training on diverse Arabic social media

[1] https://huggingface.co/UBC-NLP/MARBERTv2

text, making it suitable for general dialectal Arabic.

- **SaudiBERT (faisalq/SaudiBERT)[2]:** A model specialized in the Saudi dialect.
- **DarijaBERT (SI2M-Lab/DarijaBERT)[3]:** Designed for North African Arabic dialects, particularly Moroccan Darija.

Each model was employed with its corresponding **AutoModelForSequenceClassification** head for the sentiment classification task.

4.3 Training Strategy

Each of the three base models was fine-tuned independently on the AHaSIS training dataset using a 5-fold stratified cross-validation strategy. This means the training data was divided into five folds, and for each fold, a model was trained on four folds and validated on the held-out fold. This process was repeated for each of the three base models.

A custom Hugging Face Trainer class was utilized for the fine-tuning process. The key training parameters and enhancements were:

- **Optimizer:** AdamW optimizer.
- **Loss Function:** nn.CrossEntropyLoss with label smoothing applied at a factor of 0.1. Label smoothing helps to regularize the model and prevent overconfidence.
- **Learning Rate:** 2e-5, with a learning rate warmup schedule.
- **Batch Size:** 8 (for training, constrained by GPU memory).
- **Epochs:** Up to 6 epochs per fold, with early stopping based on validation F1-score to prevent overfitting and save the best model checkpoint.
- **FP16 Mixed-Precision Training:** Enabled to accelerate training and reduce memory consumption.
- **Hardware:** Experiments were conducted on Google Colab (single GPU).

This training regimen was repeated for each of the 5 folds for all three models (MARBERTv2, SaudiBERT, and DarijaBERT).

4.4 Ensemble Strategy

Our ensemble strategy focused on combining the predictions (logits) from the fine-tuned base models.

Validation Phase (within Cross-Validation): During the 5-fold cross-validation on the training set, for each validation fold, we obtained logits from each of the three models (MARBERTv2, SaudiBERT, DarijaBERT) trained on the other four folds. To evaluate an intermediate ensemble performance during development, a soft-voting ensemble was considered: the logits from the three models for the validation set samples were averaged, and the class with the highest average logit was chosen as the prediction. This allowed for monitoring the ensemble's potential during the cross-validation process.

Final Test Set Prediction: To generate the final predictions for the official AHaSIS test set, the following procedure was used:

1. **Per-Model Averaging Across Folds:** For each base model (MARBERTv2, SaudiBERT, DarijaBERT), the predictions (logits) on the test set were generated by each of the 5 fine-tuned instances of that model (one from each fold of the cross-validation). These 5 sets of logits for the test set were then averaged to get a single, more stable set of logits for each of the three base models.
2. **Cross-Model Averaging:** The averaged logits for the test set from MARBERTv2, SaudiBERT, and DarijaBERT were then further averaged together.
3. **Final Label Determination:** The final sentiment label for each test instance was determined by applying an argmax function to these final averaged logits.

[2] https://huggingface.co/faisalq/SaudiBERT

[3] https://huggingface.co/SI2M-Lab/DarijaBERT

The class corresponding to the highest logit value was chosen as the predicted sentiment.

This multi-stage averaging ensemble aims to combine the strengths of the diverse models and smooth out variations from individual training runs, leading to a more robust final prediction.

5 Results

Our IWAN-NLP system was evaluated on both the training data (via cross-validation) and the official AHaSIS 2025 test set. The primary evaluation metric was the macro F1-score.

Training Set Performance (5-Fold Cross-Validation): Across the 5-fold stratified cross-validation on the training set, our ensemble approach (evaluating out-of-sample predictions from each fold) yielded the following average performance:

- Cross-Validated Macro F1-score: 0.8513
- Cross-Validated Accuracy: 0.8628

These results on the training data indicated strong performance and good generalization before evaluating on the unseen test set.

Official Test Set Performance: Our final system, employing the described averaging ensemble strategy (averaging logits across folds per model, then averaging these across models for the test set), achieved the highest macro F1-score of 81%, which was the top-ranking performance among all participating systems in the shared task. This result confirms the effectiveness of our ensemble approach in leveraging the diverse strengths of the chosen Arabic language models and the robustness introduced by cross-validation and prediction averaging.

6 Error Analysis

Upon qualitative examination of the predictions made by our ensemble system, a notable pattern of confusion emerged, primarily between the neutral and negative sentiment classes. Several factors could contribute to this observation:

1. **Ambiguity of Neutrality**: Neutral sentiment itself can be inherently ambiguous. Reviews classified as neutral might contain subtle negative undertones or vice-versa, making it challenging for the models to draw a clear distinction, especially with the minimalist preprocessing approach.
2. **Sarcasm and Irony**: Social media text (and even formal reviews) can contain sarcasm and irony, where the literal meaning of words contradicts the intended sentiment. Our system did not explicitly model sarcasm detection, which is a complex NLP task in itself. Sarcastic reviews expressing negative sentiment in a seemingly positive or neutral way (or vice-versa) could easily be misclassified.
3. **Dialectal Nuances**: While we employed dialect-specific models, the interplay of different dialects within a single review or the presence of code-switching (mixing MSA with dialects, or dialects with English) could still pose challenges. Certain dialectal expressions might carry sentiment connotations that are not universally captured even by specialized models, leading to confusion, particularly with neutral or subtly negative statements.
4. **Implicit Sentiment**: Some reviews might express sentiment implicitly rather than through explicit sentiment-bearing words. For example, a factual statement about a service failure could imply negative sentiment without using overtly negative language. Detecting such implicit sentiment requires a deeper level of contextual understanding.

5. **Dataset Artifacts**: The nature of the annotation process or inherent biases within the dataset could also contribute to certain types of systematic errors. For instance, if borderline cases were predominantly labeled as neutral, the model might struggle to differentiate them from slightly negative instances. Furthermore, it was observed that some texts labeled as Darija might not be pure Darija or could be mixed with other dialects/MSA, potentially impacting the performance of the Darija-specific model on these instances and contributing to confusion.

Future work could aim to address these issues. Integrating a dedicated sarcasm detection module could be beneficial. More advanced preprocessing techniques, or alternatively, models even more robust to noisy text, might help. Furthermore, exploring multi-task learning frameworks, where the system is jointly trained to detect sentiment and other related linguistic phenomena (like sarcasm or emotion), could lead to improved performance and a better understanding of nuanced expressions.

7 Conclusion

In this paper, we described the iWAN-NLP system developed for the AHaSIS 2025 shared (Alharbi, Chafik, et al., 2025). Our system utilized an averaging ensemble of three pre-trained transformer models: MARBERTv2, SaudiBERT, and DarijaBERT. Each base model was fine-tuned using 5-fold stratified cross-validation with a custom Hugging Face Trainer, incorporating label smoothing (0.1), early stopping, FP16 mixed-precision training, and learning rate warmup. The final ensemble, derived by averaging logits across folds and models, achieved a macro F1-score of 81.0% on the official test set, and a cross-validated macro F1 of 0.8513 on the training set. This performance distinguished our system, leading to its recognition as the first-place submission in the AHaSIS 2025 shared task.

Our results demonstrate the advantages of leveraging a diverse set of language models, including those adapted to specific regional dialects, within a robust ensemble framework. The meticulous cross-validation and prediction averaging strategy proved effective in combining the strengths of these individual models and enhancing overall performance. The inclusion of modern training best practices further contributed to the stability and generalization capabilities of our system.

Future research directions could involve exploring more sophisticated ensemble techniques, investigating advanced data augmentation tailored for Arabic dialects, and incorporating explicit mechanisms for handling linguistic phenomena such as sarcasm and implicit sentiment. Our participation in AHaSIS 2025 highlights the potential of combining specialized NLP models and principled ensembling to tackle the complexities of sentiment analysis in the diverse Arabic linguistic landscape.

References

Abdul-Mageed, M., Elmadany, A., & Nagoudi, E. M. B. (2021). ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *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)* (pp. 7088–7105). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.551

Abdul-Mageed, M., Kuebler, S., & Diab, M. (2012). SAMAR: A System for Subjectivity and Sentiment Analysis of Arabic Social Media. In A. Balahur, A. Montoyo, P. M. Barco, & E. Boldrini (Eds.), *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis* (pp. 19–28). Association for Computational Linguistics. https://aclanthology.org/W12-3705/

Al Shamsi, A. A., & Abdallah, S. (2023). Ensemble Stacking Model for Sentiment Analysis of Emirati and Arabic Dialects. *Journal of King Saud University - Computer and Information Sciences*, *35*(8), 101691. https://doi.org/10.1016/j.jksuci.2023.101691

Alharbi, M., Chafik, S., Ezzini, S., Mitkov, R., Ranasinghe, T., & Hettiarachchi, H. (2025). AHaSIS: Shared Task on Sentiment Analysis for Arabic Dialects. *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Alharbi, M., Ezzini, S., Hettiarachchi, H., Ranasinghe, T., & Mitkov, R. (2025). Evaluating Large Language Models on Arabic Dialect Sentiment Analysis. *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Antoun, W., Baly, F., & Hajj, H. (2020). AraBERT: Transformer-based Model for Arabic Language Understanding. *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, 9–15. https://www.aclweb.org/anthology/2020.osact-1.2

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *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)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

El-Halees, A. (2011, December 11). *Arabic Opinion Mining Using Combined Classification Approach*. https://www.semanticscholar.org/paper/Arabic-Opinion-Mining-Using-Combined-Classification-El-Halees/4bc8728992fcf5b26ae80286b9524ff115e7d329

Gaanoun, K., Naira, A. M., Allak, A., & Benelallam, I. (2024). DarijaBERT: A step forward in NLP for the written Moroccan dialect. *International Journal of Data Science and Analytics*. https://doi.org/10.1007/s41060-023-00498-2

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., & He, K. (2018). *Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour* (arXiv:1706.02677). arXiv. https://doi.org/10.48550/arXiv.1706.02677

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., & Wu, H. (2018). *Mixed Precision Training* (arXiv:1710.03740). arXiv. https://doi.org/10.48550/arXiv.1710.03740

Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help? *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper_files/paper/2019/hash/f1748d6b0fd9d439f71450117eba2725-Abstract.html

Qarah, F. (2024). *SaudiBERT: A Large Language Model Pretrained on Saudi Dialect Corpora* (arXiv:2405.06239). arXiv. https://doi.org/10.48550/arXiv.2405.06239

Salah, Z., Al-Ghuwairi, A.-R. F., Baarah, A., Aloqaily, A., Qadoumi, B., Alhayek, M., & Alhijawi, B. (2019). A systematic review on opinion mining and sentiment analysis in social media. *International Journal of Business Information Systems*, *31*(4), 530–554. https://doi.org/10.1504/IJBIS.2019.101585

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. In Q. Liu & D. Schlangen (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Fine-tuning AraBert model for arabic sentiment detection

Mustapha Jaballah
University of Tunis,
ENSIT SIME Laboratory
musjmusj4@gmail.com

Dhaou Ghoul
HF-Lab, HIGHSYS
dhaou.ghoul@gmail.com

Ammar Mars
University of Tunis,
ISG Smart Lab Laboratory
ammar.mars@gmail.com

Abstract

Arabic exhibits a rich and intricate linguistic landscape, with Modern Standard Arabic (MSA) serving as the formal written and spoken medium, alongside a wide variety of regional dialects used in everyday communication. These dialects vary considerably in syntax, vocabulary, phonology, and meaning, presenting significant challenges for natural language processing (NLP). The complexity is particularly pronounced in sentiment analysis, where emotional expressions and idiomatic phrases differ markedly across regions, hindering consistent and accurate sentiment detection. This paper describes our submission to the Ahasis Shared Task: A Benchmark for Arabic Sentiment Analysis in the hospitality domain. This shared task focuses on advancing sentiment analysis techniques for Arabic dialects in the hotel domain. Our proposed approach achieved an F1 score of 0.88 % on the internal test set (split from the original training data), and 79.16% on the official hidden test set of the shared task. This performance secured our team second place in the Ahasis Shared Task.

1 Introduction

Sentiment analysis has become a crucial task in natural language processing (NLP), enab ling businesses and organizations to extract valuable insights from user-generated content. While significant progress has been made in sentiment analysis for English and other major languages, Arabic sentiment analysis presents unique challenges due to the language's morphological complexity, and dialectal variations. Modern Standard Arabic (MSA) coexists with numerous regional dialects that differ substantially in vocabulary, syntax, and semantics, making unified sentiment analysis particularly difficult.

The hospitality industry stands to benefit greatly from accurate sentiment analysis, as customer reviews and feedback directly impact business decisions and service quality. However, Arabic sentiment analysis in this domain faces additional difficulties, such as the prevalence of colloquial expressions that carry strong sentiment but may not appear in standard jargon.

Recent advances in transformer-based language models like BERT have shown promising results for Arabic NLP tasks. However, their application to dialectal Arabic sentiment analysis, especially in domain-specific contexts like hospitality, remains underexplored (Antoun et al., 2020). The Ahasis Shared Task provides an important benchmark for evaluating such approaches, featuring annotated hotel reviews in multiple Arabic dialects with sentiment labels (Alharbi et al., 2025a).

In this paper, we present our fine-tuned AraBERT model for Arabic sentiment detection in the hospitality domain. Our approach addresses the following key challenges :

- Handling morphological richness in Arabic text
- Adapting a pre-trained language model to domain-specific sentiment analysis

Our system achieved competitive performance in the Ahasis Shared Task, ranking second with an F1-score of 79.16%. The results demonstrate the effectiveness of transformer-based models for Arabic sentiment analysis while highlighting areas for future improvement, particularly in handling dialectal diversity and domain adaptation. The remainder of this paper is organized as follows: Section 2 reviews related work in Arabic sentiment analysis, Section 3 details our methodology, Section 4 presents and discusses our results, and Section 5 concludes with directions for future research.

2 State of the art

Sentiment analysis, the computational study of opinions and emotions in text, has evolved significantly with advancements in artificial intelligence

Proceedings of the Shared Task on Sentiment Analysis for Arabic Dialects (AHaSIS) associated with RANLP 2025, pages 14–23, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-109-7-003

(AI). For Arabic language processing, especially in multi-dialect settings, the choice of method depends on factors like data availability, dialect diversity, computational resources, and desired accuracy. This classification organizes sentiment analysis techniques into three key paradigms:

- Traditional Methods: Rule-based and classical machine learning approaches that rely on features and lexicons. These are interpretable but struggle with dialectal variations and context.
- Deep Learning Methods: Neural network-based models that automatically learn features from text, improving performance on complex language patterns (e.g., LSTMs, CNNs, and early Transformers).
- LLM-Based Methods: large language models, which leverage massive pre-trained networks for highly accurate, context-aware sentiment analysis, even in low-resource dialects (like BERT and GPT).

2.1 Traditional Methods (Rule-Based & Machine Learning)

2.1.1 Lexicon-Based (Rule-Based)

The lexicon-based approach aggregates the sentiment scores of all the words in text using a pre-prepared sentiment lexicon to assess. In this regard, in (Mataoui et al., 2018), the authors proposed syntax-based aspect detection approach for sentiments analysis in Arabic reviews. In (Elnagar et al., 2018), authors implement a polarity lexicon-based sentiment analyzer to analyze sentiment for HARD (Hotel Arabic-Reviews Dataset) dataset. A lexicon approach proposed in (Abdul-Mageed et al., 2012) reduce data sparseness through multiple morphological features, such as part of speech tagging, in addition to multiple standard features, including a polarity lexicon that handles subjectivity classification. Another approach mentioned in (Mars et al., 2015) uses traditional methods and supports huge data. It implements a MapReduce architecture based on lexicon method for sentiment analysis from Twitter.

2.1.2 Machine Learning (Classical Models)

Many works use SVM, Naïve Bayes and Random Forest combined with feature engineering (TF-IDF, n-grams) in sentiment analysis task in Arabic language. An approach cited in (Mars et al., 2017) proposes a new ontological approach based on SVM to extract sentiements from twitter. This approach uses an svm algorithm enhanced by an antology of positive and negative words.

(Akaichi, 2013) uses acronyms, interjections, and emoticons as lexicon features, as well as N-gram along with SVM algorithm. In this context, the work in (Alowaidi et al., 2017) used various classifiers such as naïve Bayes (NB) and SVM, and WordNet to extract concept features from dataset of 826 tweets. Machine Learning approaches can be improved by training the model on a large number of examples, unlike the lexicon-based approaches. It has been widely used in SA for Arabic language.

Many works implement a feature extractor based on prediction with Word2vec (Le and Mikolov, 2014), (Altowayan and Tao, 2016) and (Baly et al., 2017). In the last sudy, the accuracy reached 60.6% when using Lexicon Feature(LF) with SVM on an Egyptian dataset of 1200 tweets, while using Word2vec as a prediction-based embedding (PBE) technique along with the DL algorithm, the accuracy reached 70%. (Dhaou and Lejeune, 2020) present an ensemble classifier relying on word and character-level features developed for the Shared Task on Sarcasm and Sentiment Detection in Arabic. In this work, the F1-score reached 65.06%.

2.2 Deep Learning Methods (Neural Networks)

Deep Learning (DL) methods for sentiment analysis are a subset of machine learning techniques that use artificial neural networks, particularly deep neural networks (DNNs), to automatically learn hierarchical representations of text data.

2.2.1 Word Embeddings with Neural Networks

In (Adouane et al., 2020), authors trained BiLSTM architecture along with fastText word embedding reaching 66.78% in accuracy. In the work published in (MIHI et al., 2020), the experimentations achieved for the 4-way classification 56.3% in accuracy using Term frequency–Inverse document frequency (TF-IDF) and LR, when using the Bag of Words technique (BOW) and Support Vector Classifier (SVC), the accuracy attained 55.6%.

In the same context, (Alayba et al., 2018) presents Combined CNN and LSTM Model for Arabic Sentiment Analysis, which investigate the benefits of integrating CNNs and LSTMs and report obtained improved accuracy for Arabic sentiment analysis on different datasets. Other work evaluated several deep learning architectures using CNN and LSTM with adopting the Word2vec for vectorizing text (Al-Azani and El-Alfy, 2017). A result published in (Abbes et al., 2017) of deep learning (DL) approach for Sentiment Analysis showed that RNN outperforms DNN in term of precision. (Mars et al., 2024) proposes a method which combines different classifiers using the voting method and achieves significant F1-score value equal to 0.7027.

2.2.2 Transformer Models (Pre-LLM Era)

The Transformer architecture, introduced in 2017, revolutionized natural language processing (NLP) by replacing traditional recurrent and convolutional neural networks with self-attention mechanisms.

A new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers introduced in (Devlin et al., 2019a). It is designed to pre-train deep bidirectional representations from unlabeled text. Next, BERT pre-trained for the Arabic language in (Antoun et al., 2020) which achieved state-of-the-art performance on most tested Arabic NLP tasks.

In the same context, AraGPT2 (Antoun et al., 2021) is developed and trained on a large Arabic corpus. The results show success on different tasks including synthetic new generation, and zero-shot question answering. In addition, a framework was introduced in (Radford and Narasimhan, 2018) to achieve strong natural language understanding with a single task-agnostic model through generative pre-training and discriminative fine-tuning.

(Ghoul et al., 2024) address the challenge of Arabic sentiment analysis in short texts, where high-quality training data is often scarce. They propose three machine learning models for classifying Arabic tweets: a Voting Ensemble combining character- and word-level features, an AraBERT model with Farasa preprocessing, and a hybrid approach integrating both methods. Their best-performing model achieves a 73.98% F-score, demonstrating improvement over prior work. The study offers valuable insights for future Arabic NLP applications and services.

2.3 LLM-Based Methods (Modern Large Language Models)

Large Language Models (LLMs) represent the new edge of natural language processing (NLP), revolutionizing sentiment analysis through their deep contextual understanding, multilingual capabilities, and zero-shot learning potential.

(Miah et al., 2024) proposes an ensemble model of transformers and a large language model (LLM) that leverages sentiment analysis of foreign languages by translating them into a base language, English. The sentiment analysis task used an ensemble of pre-trained sentiment analysis models: Twitter-Roberta-Base-Sentiment-Latest, bert-base-multilingual-uncased-sentiment, and GPT-3, which is an LLM from OpenAI.

The work published in (Huang et al., 2024) proposes a solution, named AceGPT, that includes further pre-training with Arabic texts, Supervised Fine-Tuning (SFT) using native Arabic instructions, and GPT-4 responses in Arabic, alongside Reinforcement Learning with AI Feedback (RLAIF) employing a reward model attuned to local culture and values.

(Seelawi et al., 2021) propose the Arabic Language Understanding Evaluation Benchmark (ALUE), which AceGPT achieves the second best in terms of average scores for all tasks. An evaluation of ChatGPT and Bard Ai on Arabic Sentiment analysis is published in (Al-Thubaity et al., 2023). It conducts three LLMs for Dialectal Arabic Sentiment Analysis, namely ChatGPT based on GPT-3.5 and GPT-4, and Bard AI. The experiments show that GPT-4 outperforms GPT-3.5 and Bard AI in sentiment analysis classification, competing the top-performing fully supervised BERT-based language model.

Other research efforts made to evaluate the ability of LLMs for Arabic sentiment analysis which focus on single language models like AraT5 (Elmadany et al., 2022) or multiple models like (Kadaoui et al., 2023). The former introduced three powerful Arabic-specific text to-text Transformer models trained on large Modern Standard Arabic (MSA) and/or Arabic dialectal data. The latter conducted to evaluate both of Bard AI and ChatGPT LLMs for Arabic Sentiment Analysis.

3 Methodology

The methodology adopted in this study is stratified in several steps as illustrated in Figure 1.Each step is designed to fine-tune AraBERT model for Arabic sentiment detection. The following subsections detail each step.

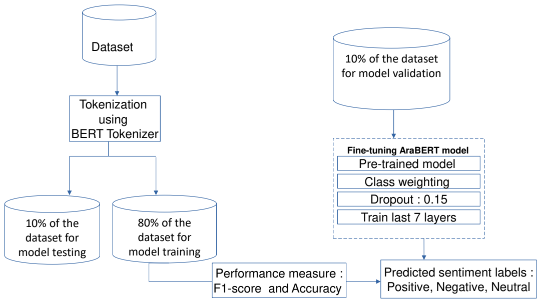


Figure 1: The system architecture.

3.1 Data

In this study, we used the Ahasis dataset which is an Arabic dataset designed for target-specific sentiment analysis. It contains a total of 860 annotated tweets related to the hospitality topic and categorized in 2 dialects Darija and Saudi as shown in Table 1.

Dialect	Darija	Saudi	Total
Negative	168	168	336
Neutral	108	108	216
Positive	154	154	308
Total	430	430	860

Table 1: Distribution of sentiment and dialect in the dataset.

This dataset serves as a benchmark for the tasks of sentiment analysis and offers valuable opportunities for exploring the interaction between different dimensions of opinion and evaluating learning models.

The Table 2 illustrates some examples from the dataset.

3.2 Tokenization

The raw text is tokenized using a BERT tokenizer (Devlin et al., 2019b). This step is crucial. It converts the text into a format adequate for input into the BERT model (Antoun et al.). The tokenization process involves setting a maximum sequence length of 128 tokens, with padding and truncation applied to maintain uniform input sizes. This step is crucial, as BERT requires fixed-length inputs for efficient batch processing. Additionally, the tokenizer performs subword tokenization, which is especially beneficial for Arabic given its complex morphology (Abadi et al., 2015).

3.3 AraBERT Fine-tuning

The core of the methodology centers on fine-tuning a pre-trained BERT model using the Arabic text dataset Ahasis. Specifically, we used the aubmindlab / bert-base-arabertv02 Twitter model (Antoun et al.), which is trained on Arabic Twitter data, making it particularly effective for social media text classification.

To address class imbalance, we implement class weighting, assigning higher weights to the "neutral" class (Hinton et al., 2012). Indeed, we use two approaches. First, we automatically calculate weights for each sentiment class based on how frequently they appear - giving more importance to rare sentiments and less to common ones. We then go a step further by doubling the weight for the particularly underrepresented 'neutral' class to make sure those examples aren't overlooked. Second, we customize the training process to use these weights in error calculations, so when the model makes mistakes on less common sentiments, those errors count more heavily in the learning process.

This combination helps balance the model's attention across all sentiment categories, preventing it from favoring only the most common ones.

A dropout rate of 0.15 is applied to the hidden layers and the classifier to reduce the risk of overfitting (Hinton et al., 2012). Moreover, only the last seven layers of the BERT model are fine-tuned, allowing the model to adapt to the specific task while preserving the general language understanding learned during pretraining (Kumar et al., 2021).

The key hyperparameter settings are summarized in Table 3, providing a clear overview of the model configuration. Once fine-tuned, the model is used to predict sentiment labels for previously unseen Arabic text. The output is a set of predicted labels corresponding to the input, showcasing the model's practical utility in real-world applications such as social media monitoring and sentiment analysis. Overall, this methodology ensures a robust, systematic approach to Arabic text classification using AraBERT, with an emphasis on performance and generalization.

4 Results and discussion

The results obtained by our model for the test and dev set are presented in Table 4 The confusion matrix Figure 2 visualizes the performance of the model.The model performs very well on the 'negative' class with 100% precision and recall (35

Text	Dialect	Sentiment
هذا أسوأ فندق حجزت فيه ليلة و لكني ما قدرت اجلس فيه الا ليلة وحده بس (This is the worst hotel I've ever booked for a night, but I couldn't even stay there for more than one night)	saudi	negative
الشاطئ ممتاز لكن ماهو نظيف (The beach is excellent, but it's not clean.)	saudi	Neutral
فندق خايب بزاف، هاد الفندق من أسوأ الفنادق اللي جربتهم. كيقولوا باللي هو نجوم، ولكن ما كيستاهلحتى نجمة وحدة(The hotel is very bad, this is one of the worst hotels I've ever stayed in. They claim it's star-rated, but it doesn't even deserve one star.)	darija	Negative
انا نزلت في هذا الفندق مرتين و كلها كانت مريحة(I stayed at this hotel twice, and both times were comfortable.)	saudi	positive
كان كلشي مزيان، خاصة الغرف اللي كايطلعو على الكعبة (Everything was nice, especially the rooms that overlook the Kaaba.)	darija	positive
الفطور كان معقول، ما جربتش شي وجبات اخرى، و الموظفين كانو مزيان بزاف (The breakfast was reasonable, I didn't try any other meals, and the staff were very nice.)	darija	neutral

Table 2: Examples of annotated tweets

Hyperparameter	Value
Learning Rate	1.1e-4
Batch Size	64
Weight Decay	0.15
Number of Frozen Layers	5
Warmup Ratio	0.25
Dropout Rate	0.15
Maximum Sequence Length	128
Training Epochs	20
Gradient Accumulation Steps	2
Learning Rate Scheduler	Cosine

Table 3: Optimal hyperparameters for AraBERT fine-tuning.

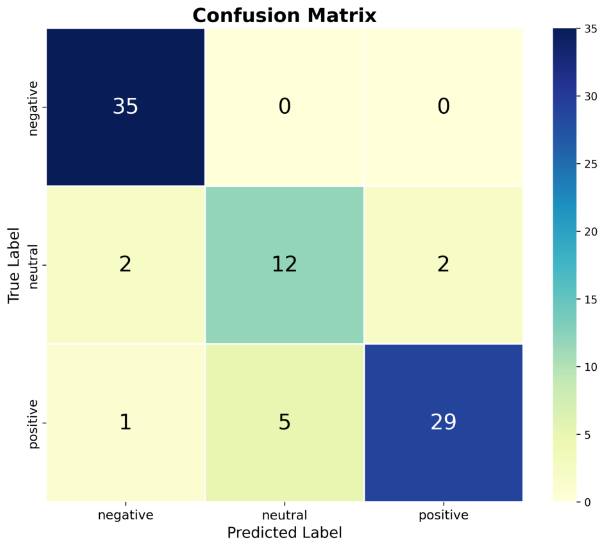


Figure 2: Confusion matrix

correctly predicted, 0 errors).Most errors occur between 'neutral' and 'positive', especially confusing positive as neutral, and Minor confusion of these two classes with negative.

As shown in the Figure 3a, the model starts with high loss (1.22) but quickly improves. By epoch 8, the loss drops to 0.34 and stays around 0.25-0.30 for the rest of training. The small rise at epoch 13 suggests the model might be starting to overfit. The best result happens at epoch 11 (loss of 0.256), which would be a good place to stop training.

Both accuracy and F1-score improve together, going from very low (under 30%) to very good (over 90%) by epoch 8 as shown in the Figure 3b and Figure 3c. They reach their highest point at epoch 11 (93%), then stay about the same or drop slightly. This shows the model learns well at first, but stops getting better after epoch 11.

The model works well and learns quickly in the first 11 epochs. After that, it doesn't improve much. To save time and get the best results, we could stop training at epoch 11. To make the model even better, we might need to add more training data, as

Class	Precision	Recall	F1-Score	Support
negative	0.92	1.00	0.96	35
neutral	0.71	0.75	0.73	16
positive	0.74	0.83	0.78	35
Accuracy		0.88		86
Macro	0.85	0.86	0.85	86
weighted	0.89	0.88	0.88	86

Table 4: Classification report for the fine-tuned AraBERT model

more training epochs won't help much. This work is cited in (Alharbi et al., 2025b) which summarizes all the Ahasis shared task participants' works.

4.1 Discussion

After reviewing the results and exploring the dev set in more detail, we discovered wrongly labeled examples as shown in Table 5. In the comparison between predicted and true labels, several misclassifications highlight potential areas for model improvement. The model frequently misclassified positive reviews as neutral, particularly when the language was nuanced or mixed (e.g., "الانترنت جيد مع انه قوي بس في الاماكن القريبة من الراوتر").

This suggests the model may struggle with contextual understanding or assigning higher confidence to neutral predictions when sentiment is subtly expressed. Additionally, the model incorrectly labeled a sarcastic positive review ("فندق للناس اللي ما كايكميوش") as negative, indicating difficulty in detecting irony or sarcasm. The high neutral probabilities (e.g., 0.99, 0.96) in cases where the true label was positive suggest an over-reliance on neutral classifications, possibly due to imbalanced training data or insufficient sensitivity to positive sentiment cues. Further refinement, such as incorporating sarcasm detection or rebalancing class weights, could enhance performance.

Our model demonstrates strong but variable performance across different evaluation sets. During training, it achieved its peak F1-score of 0.93 (93%) on the validation data by epoch 11. However, testing on unseen datasets revealed notable performance discrepancies, highlighting key considerations for real-world deployment.

The model attained an F1-score of 0.88 on the standard test set, reflecting a modest 5% decline from the validation score (0.93). This marginal drop is consistent with typical generalization behavior, suggesting that the model performs robustly on data sampled from a similar distribution as the training set. However, the slight discrepancy may indicate minor overfitting to the validation data or subtle differences in data partitioning.

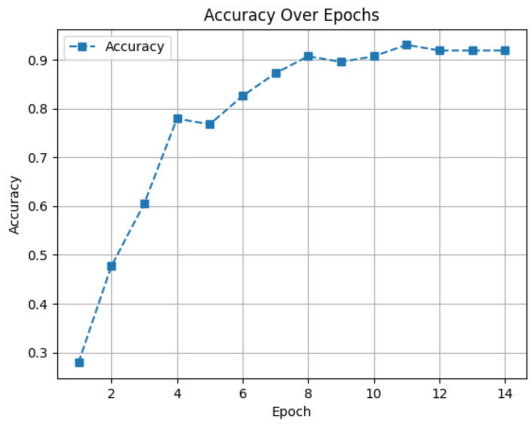
A more substantial performance degradation was observed on the blind test set, where the F1-score dropped to 0.79 a 9% decrease compared to the main test set. This discrepancy suggests:

- Distributional differences between the blind test data and the training/validation sets, possibly due to unseen variations or domain shifts.
- Limited generalizability of some learned patterns, implying that the model may rely on features that do not transfer effectively to entirely new data.
- Potential biases in the original dataset, where certain underrepresented scenarios were not adequately captured during training.

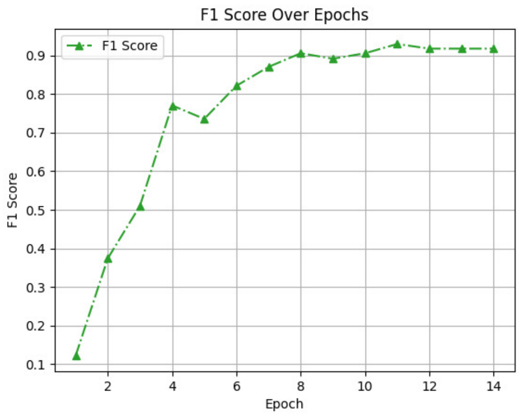
Overall, the model exhibits promising performance but suffers a 12–14% reduction in F1-score (from 0.93 to 0.79–0.88) when evaluated on unseen data. The blind test results underscore the importance of assessing models beyond standard test sets, as they reveal critical gaps in generalization that conventional evaluations may overlook. To enhance model robustness, future work should consider:

- Data augmentation and domain adaptation to improve generalization across diverse scenarios.
- Expanded dataset collection, particularly targeting underrepresented or edge cases to reduce distributional biases.
- Further analysis of feature representations to identify and mitigate non-transferable learned patterns.

(a) Progression of Validation Loss During Model Training



(b) Evolution of Model Accuracy with Increasing Training Epochs



(c) F1 Score Across Epochs

Figure 3: Training Metrics: Validation Loss and F1 Score Across Epochs.

Moreover, the dataset contains labeling inconsistencies and dialectal variations that may negatively impact model performance. For instance, some ground-truth labels appear questionable, such as labeling "بوفيه الفطور مزيان" (the breakfast buffet is good) as neutral rather than positive, or classifying a critical statement about a cramped room ("الغرفة كانت مزحمة") as positive.

These inconsistencies suggest possible annotation errors or subjective biases in the dataset. Such inaccuracies can mislead the model during training, causing it to learn incorrect sentiment associations and reducing its generalization capability. To improve reliability, a thorough review of the labels, particularly ambiguous terms and borderline cases, should be conducted, possibly with dialect-specific guidelines to ensure consistency. Otherwise, the model may propagate these errors in its predictions, particularly in sentiment analysis tasks where contextual and cultural nuances play a key role.

While the current results are encouraging, the blind test performance highlights the need for improvements in handling novel data, which is crucial for real-world applicability.

5 Conclusion and future works

In this paper, we suggest an approach on multi-dialect sentiment detection in hotel reviews. To validate the effectiveness of our approach, we used Ahasis dataset which consists of Arabic text samples labeled with sentiment and dialect. The findings from the experimentation confirm that our proposed method attains an F1-score of 0.79, indicating its performance compared to baseline models.

The success of the proposed approach suggests that leveraging multi-dialect datasets like Ahasis can improve model robustness. However, future research should explore deeper dialectal nuances, including code-switching between MSA and dialects, to enhance accuracy further.

Text	Dialect	Train label	Corrected label
بوفيه الفطور مزيان The breakfast buffet is good	Darija	neutral	positive
واحد الشي اللي يمكن نقول بيه هو، الغرفة كانت مزحمة شوية، كانت غرفة ثلاثية، إلا بلي حطوا شي سرير إضافي لي يصير غرفة رباعية One thing worth mentioning is that the room was a bit cramped. It was a triple room, but they added an extra bed to make it a quadruple room	Darija	positive	negative
كان منعش وطازج وحو حلو، خيارات حلوه من الاكل وخيارات كثيرة للمتعه بالعطلات قضينا وقت روعه في حفل الزواج واللي كان استثنائي وفخم It was refreshing, fresh, and very sweet. There were delicious food options and plenty of choices for holiday enjoyment. We had an amazing time at the wedding venue, which was exceptional and luxurious	Saudi	neutral	positive

Table 5: Examples of wrongly annotated train tweets

Moreover, to further improve sentiment analysis performance, particularly for darija, a promising direction is the integration of lexicon-based sentiment analysis (also known as dictionary-based sentiment analysis). This approach involves:

- Constructing a Domain-Specific Sentiment Lexicon.
- Developing a curated list of darija words and phrases annotated with sentiment polarity (positive, negative, neutral)
- Addressing dialectal variations and contextual ambiguities (e.g., words whose polarity shifts across regions).
- Using the lexicon to adjust classifier confidence scores, either as additional input features or as a post-processing step.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Mariem Abbes, Zied Kechaou, and Adel M. Alimi. 2017. Enhanced deep learning models for sentiment analysis in arab social media. In *Neural Information Processing*, pages 667–676, Cham. Springer International Publishing.

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.

Wafia Adouane, Samia Touileb, and Jean-Philippe Bernardy. 2020. Identifying sentiments in Algerian code-switched user-generated comments. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2698–2705, Marseille, France. European Language Resources Association.

Jalel Akaichi. 2013. Social networks' facebook' statutes updates mining for sentiment classification. In *2013 International Conference on Social Computing*, pages 886–891.

Sadam Al-Azani and El-Sayed M. El-Alfy. 2017. Hybrid deep learning for sentiment polarity determination of arabic microblogs. In *Neural Information Processing*, pages 491–500, Cham. Springer International Publishing.

Abdulmohsen Al-Thubaity, Sakhar Alkhereyf, Hanan Murayshid, Nouf Alshalawi, Maha Omirah, Raghad Alateeq, Rawabi Almutairi, Razan Alsuwailem, Manal Alhassoun, and Imaan Alkhanen. 2023. Evaluating ChatGPT and bard AI on Arabic sentiment analysis. In *Proceedings of ArabicNLP 2023*, pages 335–349, Singapore (Hybrid). Association for Computational Linguistics.

Abdulaziz M. Alayba, Vasile Palade, Matthew England, and Rahat Iqbal. 2018. A combined cnn and lstm model for arabic sentiment analysis. In *Machine Learning and Knowledge Extraction*, pages 179–191, Cham. Springer International Publishing.

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)*.

Sana Alowaidi, Mustafa Saleh, and Osama Abulnaja. 2017. Semantic sentiment analysis of arabic texts. *International Journal of Advanced Computer Science and Applications*, 8(2).

A. Aziz Altowayan and Lixin Tao. 2016. Word embeddings for arabic sentiment analysis. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 3820–3825.

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

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.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. AraGPT2: Pre-trained transformer for Arabic language generation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 196–207, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Ramy Baly, Georges El-Khoury, Rawan Moukalled, Rita Aoun, Hazem Hajj, Khaled Bashir Shaban, and Wassim El-Hajj. 2017. Comparative evaluation of sentiment analysis methods across arabic dialects. *Procedia Computer Science*, 117:266–273. Arabic Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. Bert: Pre-training of deep bidirectional transformers for language understanding.

Ghoul Dhaou and Gaël Lejeune. 2020. Comparison between voting classifier and deep learning methods for Arabic dialect identification. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 243–249, Barcelona, Spain (Online). Association for Computational Linguistics.

AbdelRahim Elmadany, Muhammad Abdul-Mageed, et al. 2022. Arat5: Text-to-text transformers for arabic language generation. In *Proceedings of the 60th annual meeting of the association for computational linguistics (Volume 1: Long papers)*, pages 628–647.

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.

Dhaou Ghoul, Jérémy Patrix, Gaël Lejeune, and Jérôme Verny. 2024. A combined arabert and voting ensemble classifier model for arabic sentiment analysis. *Natural Language Processing Journal*, 8:100100.

Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *ArXiv*, abs/1207.0580.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, Ziche Liu, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. AceGPT, localizing large language models in Arabic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8139–8163, Mexico City, Mexico. Association for Computational Linguistics.

Karima Kadaoui, Samar Magdy, Abdul Waheed, Md Tawkat Islam Khondaker, Ahmed El-Shangiti, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. TARJAMAT: Evaluation of bard and ChatGPT on machine translation of ten Arabic varieties. In *Proceedings of ArabicNLP 2023*, pages 52–75, Singapore (Hybrid). Association for Computational Linguistics.

Sachin Kumar, Shuly Wintner, Noah A. Smith, and Yulia Tsvetkov. 2021. Topics to avoid: Demoting latent confounds in text classification.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Bejing, China. PMLR.

Ammar Mars, Mohamed Salah Gouider, and Lamjed Ben Saïd. 2015. A new big data framework for customer opinions polarity extraction. In *International Conference: Beyond Databases, Architectures and Structures*, pages 518–531. Springer.

Ammar Mars, Sihem Hamem, and Mohamed Salah Gouider. 2017. New ontological approach for opinion polarity extraction from twitter. In *Computational Collective Intelligence: 9th International Conference, ICCCI 2017, Nicosia, Cyprus, September 27-29, 2017, Proceedings, Part II 9*, pages 448–458. Springer.

Ammar Mars, Mustapha Jaballah, and Dhaou Ghoul. 2024. Ishfmg_tun at stanceeval: Ensemble method for arabic stance evaluation system. In *Proceedings of The Second Arabic Natural Language Processing Conference, ArabicNLP 2024, Bangkok, Thailand, August 16, 2024*, pages 832–836. Association for Computational Linguistics.

M'hamed Mataoui, Tadj Eddine Bendali Hacine, Imad Tellache, Abdelghani Bakhtouchi, and Omar Zelmati. 2018. A new syntax-based aspect detection approach for sentiment analysis in arabic reviews. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–6.

Md Saef Ullah Miah, Md Mohsin Kabir, Talha Bin Sarwar, Mejdl Safran, Sultan Alfarhood, and MF Mridha. 2024. A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and llm. *Scientific Reports*, 14(1):9603.

Soukaina MIHI, Brahim AIT BEN ALI, Ismail EL BAZI, Sara AREZKI, and Nabil LAACHFOUBI. 2020. Mstd: Moroccan sentiment twitter dataset. *International Journal of Advanced Computer Science and Applications*, 11(10).

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Haitham Seelawi, Ibraheem Tuffaha, Mahmoud Gzawi, Wael Farhan, Bashar Talafha, Riham Badawi, Zyad Sober, Oday Al-Dweik, Abed Alhakim Freihat, and Hussein Al-Natsheh. 2021. ALUE: Arabic language understanding evaluation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 173–184, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Enhancing Arabic Dialectal Sentiment Analysis through Advanced Data Augmentation Techniques

Md. Rafiul Biswas
Hamad Bin Khalifa University
Doha, Qatar
mbiswas@hbku.edu.qa

Wajdi Zaghouani
Northwestern University in Qatar
Education City, Doha, Qatar
wajdi.zaghouani@northwestern.edu

Abstract

This work addresses the challenge of Arabic sentiment analysis in the hospitality domain in all dialects by using data augmentation techniques. We created a pipeline with three simple techniques: context-based paraphrasing, pattern-based sentence generation, and domain-specific word replacement. Our method preserves the original dialect features, meanings, and key classification details while adding diversity to the training data. It also includes automatic fallback between methods to handle challenges effectively. We used the Fanar API for dialectal data augmentation in the hospitality domain. The AraBERT-Large-v02 model was fine-tuned on original and augmented data, showing improved performance. This study helps solve the problem of limited dialect data in Arabic NLP and offers an effective framework that is useful for other Arabic text analysis tasks.

1 Introduction

In the Arabic-speaking world, most of the content on social networks and online reviews is written in regional dialects rather than Modern Standard Arabic (MSA). Arabic dialectal sentiment analysis aims to identify emotions (positive, negative, neutral) in social media texts written in regional Arabic dialects. These dialects, such as Egyptian, Saudi, and Moroccan Arabic, exhibit significant variations in vocabulary, grammar, and syntax, making sentiment analysis a complex task due to the lack of standard spelling and limited NLP tools (Salloum, 2021; Baly et al., 2017).

In the hospitality industry, which includes hotels, restaurants and tourism services, sentiment analysis is crucial to analyze customer feedback to improve service quality and customer satisfaction (Musanovic et al., 2021; Kim et al., 2022). Analyzing sentiments from reviews written in different Arabic dialects can provide valuable insights for companies operating in Arabic-speaking countries (Al-Thubaity et al., 2018). Unlike formal Modern Standard Arabic (MSA), these dialects are used in daily chats and reviews, but they do not follow strict rules, making it hard for computers to understand them (El-Naggar et al., 2017). In addition, there are not many tools or datasets built specifically for these dialects. Dataset for context-specific dialect like hospitality is essential to build machine learning tool. The release of a dialectal sentiment dataset and the organization of a shared task in the Hospitality Domain by the Ahasis organizing team is a pivotal contribution to the Arabic NLP community (Alharbi et al., 2025a,b). It will encourage the research community to develop NLP tool in specific domain.

2 Dataset Description

Ahasis shared task (Alharbi et al., 2025a) includes Arabic sentiment analysis texts from two dialects: Moroccan Darija and Saudi Arabic. The training set contains 860 balanced sentences, with 430 samples from each dialect. Both subsets share identical sentiment distributions: 39.07% negative, 25.12% neutral, and 35.81% positive. This balance helps prevent the model from favoring any dialect or sentiment class. The test set maintains this dialectal balance with 108 samples each, ensuring robust evaluation. See Table 1 for more details.

This dataset offers unique advantages for Arabic NLP sentiment analysis by covering dialectal diversity beyond standard Arabic, ensuring consistent sentiment annotations, and representing all sentiment classes adequately. Its dual-dialect structure supports both dialect-specific, cross-dialect experiments, industry applications like customer feedback analysis, review summarization, and reputation management.

Proceedings of the Shared Task on Sentiment Analysis for Arabic Dialects (AHaSIS) associated with RANLP 2025, pages 24–28, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-109-7-004

However, the dataset has some limitations. The relatively small size (860 training samples) may limit the model's ability to generalize to more complex patterns. It is insufficient size for training large-scale models, particularly deep neural networks, without overfitting risks. The absence of contextual topic metadata also tied to specific topics may be missed during modeling.

Table 1: Train Dataset Distribution

Dialect	Negative	Neutral	Positive	Total
Darija	168	108	154	430
Saudi	168	108	154	430

The aim of this shared task is to correctly predict the sentiment based on dialect. The test set provides dialects for each text to correctly predict sentiment. Research suggests that hybrid methods, which combine word lists with machine learning, often perform better than traditional approaches. Recent advances, such as transformer models, show promising results, but require more dialect-specific resources.

3 System Description

Transformer based model requires a good amount of data for training the model. However, the task lacks of original data. Data augmentation plays a crucial role in enhancing the performance and generalizability of text classification models, especially in low-resource scenarios (Shah et al., 2024). Our system addresses these challenges through a multi-faceted approach to text augmentation. We explained the system in the following sections.

3.1 Dataset Preprocessing

Data preprocessing is essential to preserve the semantic integrity of dialectal Arabic for optimizing the with state-of-the-art language models (e.g., AraBERT). The preprocessing pipeline involved several steps to clean and standardize the textual data. HTML markup was removed to eliminate irrelevant formatting tags. URLs, email addresses, and social media mentions were replaced with special tokens, ensuring a consistent structure across diverse inputs. To normalize the text, Arabic diacritics (tashkeel) and elongation characters (tatweel) were stripped. Whitespace was optionally inserted around punctuation and special characters to facilitate better tokenization. Emojis were retained to preserve sentiment-related cues, and extra spaces were removed to maintain clean and consistent formatting throughout the dataset.

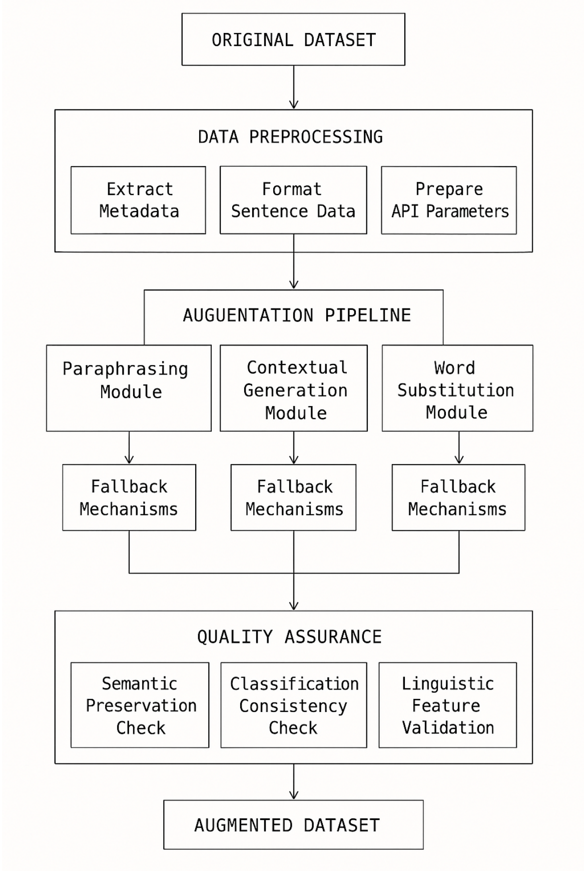


Figure 1: Data Augmentation

3.2 Data Augmentation

We have proposed an innovative approach for data augmentation using Large Language Model (LLM) to mitigate low resource challenges. Figure 1 describes the data augmentation scenario. The system is designed to preserve critical linguistic features including dialectal characteristics, domain context, and classification-relevant patterns. We performed data augmentation in three ways (Text Paraphrasing using LLM, Contextual Text Generation, and Word Substitution). Finally, we combined all the augmented data.

Text Paraphrasing using LLM: We integrated FANAR API (Team et al., 2025) to paraphrase the raw data. The paraphrasing module generates alternative versions of existing Arabic sentences while preserving their essential characteristics: preserving dialectal features, maintaining of domain context (tourism, hotel and reservations), and retenting tonal characteristics. We have created instructed prompt to instruct the FANAR model. We provided

10 example with system prompt, user prompt, text, dialect, sentiment. Finally, we asked the FANAR API to perform text paraphrasing.

Contextual Text Generation: We avoided sensitive content by focusing on language structure (e.g., hate speech, vulgarity, offensive language). We analyzed example sentences to find patterns, and then created new sentences from these patterns. If standard generation faces issues, it rebuilds sentences directly from identified patterns. This module creates new sentences matching patterns from the original dataset. It increases dataset variety while keeping original language features, generating text based on topic (hospitality), matching dialect (Saudi, Darija), and keeping important features for classification overall pattern.

Word Substitution: The module slightly modifies sentences without changing important meanings that affect classification. We replaced key words with suitable synonyms, and kept original sentence meaning and structure. We made minimal changes to avoid classification errors. We used synonyms related to the specific topic (Hospitality) ensuring sentence meaning stays consistent.

Fallback Mechanism: In all three cases mentioned above, we trained and evaluated model performance on the validation dataset. Our key innovation lies in the implementation of comprehensive fallback mechanisms (Liu et al., 2023), which ensure robustness when primary augmentation strategies face challenges. A fallback mechanism automatically activates when the preferred option is unavailable or fails. To maintain data augmentation quality, we assessed semantic preservation, classification consistency (i.e., label distribution across positive, negative, and neutral), and linguistic integrity through human evaluation. We validated the augmentation outputs manually and applied fallback procedures when necessary to improve outcomes. This process enabled us to achieve high-quality data augmentation for model training. Table 2 summarizes the total number of augmented samples.

Table 2: Data Augmentation

Dialect	Negative	Positive	Neutral
Darija	647	613	421
Saudi	617	587	394

4 Experimental Setup

Our sentiment analysis leveraged with AraBERT-Large-v02 (”aubmindlab/bert-large-arabertv02”) (Antoun et al., 2020) and CAMeLBERT mix(Inoue et al., 2021), optimized for dialect-specific Arabic preprocessing. We stratified data sampling into (80% training, 20% validation) to retain sentiment class proportions. We employed employed AdamW (learning rate: 2e-5, epsilon: 1e-8) with linear scheduling (no warmup). The experiment continued for 10 epochs, batch size 16, and maximum token length 128, alongside gradient clipping (threshold: 1.0) to stabilize training.

We utilized Google Colab Pro, NVIDIA A100 GPUs, enabling efficient model optimization with accelerated processing capabilities suitable for transformer-based architectures like AraBERT-Large-v02. This infrastructure significantly reduced training duration and facilitated rapid experimentation and hyperparameter tuning.

5 Results

Figure 2 illustrates the training and validation loss alongside validation performance metrics over 10 epochs. Training loss consistently decreased and validation loss reached stability around epoch 6, reflecting a balanced training without notable overfitting. Our fused model demonstrated consistent performance across both validation and test datasets for Arabic dialectal sentiment classification. The results (see Table 3) show a slight decrease in performance metrics when moving from validation to test data. The slight performance decrease on the test set within expected ranges and indicates good generalization capabilities.

Table 3: Validation vs. Test Performance Comparison

Metric	Validation	Test
Accuracy	0.780	0.750
Precision	0.780	0.750
Recall	0.780	0.750
F1-Score	0.780	0.750

The competition winner achieved (F1-score 0.81) where we achieved (F1-score 0.75) showing 0.6 back to winner. This indicates that while our model is competitive, further improvements in data augmentation, model tuning, or handling of dialectal variations may help bridge this gap.

Figure 2: Training loss Vs Validation Loss

6 Conclusion

In this study, we explore the analysis of Arabic dialectal sentiment using AraBERT-Large-v02, supported by domain-specific preprocessing and controlled data augmentation. Our results demonstrate that careful handling of dialectal features and balanced data splitting are crucial to achieving reliable sentiment classification. The model's ability to maintain consistent performance across Darija and Saudi dialects suggests that it could be deployed in real-world applications requiring sentiment monitoring across diverse Arabic dialectal contexts. Future work should expand to include additional Arabic dialects and explore multitask learning approaches, advanced augmentation, or domain adaptation techniques to further improve classification performance.

Acknowledgments

This study was supported by the grant NPRP14C-0916-210015, awarded by the Qatar Research, Development and Innovation Council (QRDI).

References

Abdulmohsen Al-Thubaity, Qubayl Alqahtani, and Abdulaziz Aljandal. 2018. Sentiment lexicon for sentiment analysis of saudi dialect tweets. *Procedia computer science*, 142:301–307.

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 *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Ramy Baly, Georges El-Khoury, Rawan Moukalled, Rita Aoun, Hazem Hajj, Khaled Bashir Shaban, and Wassim El-Hajj. 2017. Comparative evaluation of sentiment analysis methods across arabic dialects. *Procedia Computer Science*, 117:266–273.

Nadine El-Naggar, Yasser El-Sonbaty, and Mohamad Abou El-Nasr. 2017. Sentiment analysis of modern standard arabic and egyptian dialectal arabic tweets. In *2017 computing conference*, pages 880–887. IEEE.

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*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Taekyung Kim, Hwirim Jo, Yerin Yhee, and Chulmo Koo. 2022. Robots, artificial intelligence, and service automation (raisa) in hospitality: sentiment analysis of youtube streaming data. *Electronic Markets*, 32(1):259–275.

Ye Liu, Semih Yavuz, Rui Meng, Meghana Moorthy, Shafiq Joty, Caiming Xiong, and Yingbo Zhou. 2023. Modeling uncertainty and using post-fusion as fallback improves retrieval augmented generation with llms. *arXiv preprint arXiv:2308.12574*.

Jelena Musanovic, Raffaella Folgieri, Maja Gregoric, et al. 2021. Sentiment analysis and multimodal approach applied to social media content in hospitality industry. *TOURISM IN SOUTH EAST EUROPE...*, 6:533–544.

SA Salloum. 2021. Sentiment analysis in dialectal arabic: a systematic review. *Advanced machine learning technologies and applications: proceedings of AMLTA*.

Uzair Shah, Md Rafiul Biswas, Marco Agus, Mowafa Househ, and Wajdi Zaghouani. 2024. Mememind at araieval shared task: generative augmentation and feature fusion for multimodal propaganda detection in arabic memes through advanced language and vision models. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 467–472.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, Majd Hawasly, Mus'ab Husaini, Soon-Gyo Jung, Ji Kim Lucas, Walid Magdy, Safa Messaoud, Abubakr Mohamed, Tasnim Mohiuddin, Basel Mousi, Hamdy Mubarak, Ahmad Musleh, Zan Naeem, Mourad Ouzzani, Dorde Popovic, Amin Sadeghi, Husrev Taha Sencar, Mohammed Shinoy, Omar Sinan, Yifan Zhang, Ahmed Ali, Yassine El Kheir, Xiaosong Ma, and Chaoyi Ruan. 2025. Fanar: An arabic-centric multimodal generative ai platform.

Ahasis Shared Task: Hybrid Lexicon-Augmented AraBERT Model for Sentiment Detection in Arabic Dialects

Shimaa Amer Ibrahim
Northwestern University in Qatar
shimaa.ibrahim@northwestern.edu

Mabrouka Bessghaier
Northwestern University in Qatar
mabrouka.bessghaier@northwestern.edu

Wajdi Zaghouani
Northwestern University in Qatar
wajdi.zaghouani@northwestern.edu

Abstract

This work was conducted as part of the Ahasis@RANLP–2025 shared task, which focuses on sentiment detection in Arabic dialects within the hotel review domain. The primary objective is to advance sentiment analysis methodologies tailored to dialectal Arabic. Our work combines data augmentation with a hybrid model that integrates AraBERT and our created sentiment lexicon. Notably, our hybrid model significantly improved performance, reaching an F1-score of 0.74, compared to 0.56 when using only AraBERT. These results highlight the effectiveness of lexicon integration and augmentation strategies in enhancing both the accuracy and robustness of sentiment classification in dialectal Arabic.

1 Introduction

Arabic is characterized by a complex linguistic landscape, where Modern Standard Arabic (MSA) serves formal and written communication, while a wide range of regional dialects dominate daily spoken and informal written discourse. The coexistence of MSA and diverse regional dialects, along with the language's rich morphology, poses substantial challenges for Natural Language Processing (NLP) tasks, particularly sentiment analysis. Dialectal Arabic lacks standardized spelling, varies significantly across regions, and is underrepresented in NLP resources compared to MSA. In response to this gap, our work focuses on sentiment classification in Saudi Arabic and Darija (Moroccan Arabic) using a dataset of hotel reviews from the Ahasis Shared Task (Alharbi et al., 2025a). Each review is labeled with one of three sentiment classes: positive, neutral, or negative. To address the limitations of dialectal data and improve generalizability, we propose a hybrid sentiment classification model that combines the strengths of AraBERT with lexicon-based sentiment features. We also apply data augmentation techniques to enrich the training set and enhance performance across dialectal variations. The remainder of this paper is organized as follows: Section 2 presents related work on Arabic sentiment analysis. Section 3 provides a detailed description of the Ahasis shared task, while Section 4 outlines the proposed methodology, including data augmentation, lexicon building and integration, and model development, followed by experimental results.

2 Related work

Arabic sentiment analysis remains a significant challenge due to the coexistence of MSA and a wide range of regional dialects, which differ substantially in morphology, syntax, and vocabulary. These linguistic variations are especially problematic in informal, domain-specific contexts such as hotel reviews.

Recent works are now focusing more on dialect-aware modeling. (Abo et al., 2024) constructed a polarity lexicon for the Saudi dialect and demonstrated the benefits of dialect-specific preprocessing in improving classification performance on hotel-related data. (Obiedat et al., 2021) reviewed 21 Arabic aspect-based sentiment analysis (ABSA) studies and identified hotel reviews as a frequently used benchmark. However, they noted persistent limitations, including the scarcity of multi-dialect resources and limited use of augmentation or hybrid modeling. Recent reviews further contextualize the evolution of Arabic sentiment analysis (ASA) methodologies. (Al Katat et al., 2024) conducted a large-scale systematic review of 100 studies and confirmed the dominance of deep learning and transformer-based models in achieving high performance, especially in dialectal and informal contexts. For example, (Ghoul et al., 2024) integrated AraBERT embeddings with SVM and

Proceedings of the Shared Task on Sentiment Analysis for Arabic Dialects (AHaSIS) associated with RANLP 2025, pages 29–34, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-109-7-005

Random Forest, showing improved classification in low-resource settings. Moreover, (Firdous and Iqbal, 2025) analyzed performance trends in traditional and deep learning models. They found that AraBERT consistently outperformed classical ML algorithms, though performance declined in informal or highly dialectal texts, highlighting the need for improved annotated corpora. (Aladeemy et al., 2024) emphasized the lack of standardized, domain-adapted lexicons and advocated for the development of comprehensive tools for dialectal domains like hospitality. Additionally, (Alosaimi et al., 2024) introduced a hybrid AraBERT-LSTM architecture and showcased the benefit of combining contextual embeddings with sequential modeling for sentiment tasks.

Arabic data augmentation has also evolved across multiple techniques. Lexicon-based strategies, such as in (Duwairi and Abushaqra, 2021), used synonym replacement from Arabic WordNet to generate semantically similar variants. Embedding-based methods like (Alkadri et al., 2022) employed AraVec to substitute words based on cosine similarity, expanding lexical diversity. Besides, back-translation has been applied to generate synthetic paraphrase corpora (Al-shameri and Al-Khalifa, 2024), while generative models such as AraGPT2 have been used to create augmented examples for minority sentiment classes (Abdhood et al., 2025). Among generative approaches, AraT5 stands out as a powerful text-to-text model tailored for Arabic. Introduced by (Bani-Almarjeh and Kurdy, 2023) and further evaluated in (Nagoudi et al., 2022; Masri et al., 2025). AraT5 has demonstrated strong performance in summarization and paraphrasing tasks, making it a valuable tool for data augmentation in dialect-sensitive NLP applications. Building on these foundations, our system applies AraT5-based paraphrasing to diversify training data, particularly to address the limited sample size of the shared task dataset. We also constructed a custom sentiment lexicon from the dataset, incorporating dialectal stopword expansion and frequency-based scoring. Together, these augmentation and lexicon strategies strengthen sentiment modeling across dialects in the hotel review domain.

3 Task Description

The Ahasis task [1] focuses on sentiment analysis on Arabic dialects in the hospitality domain. Participants should classify sentiment as positive, neutral, or negative across different Arabic dialects (i.e Saudi and Moroccan). This dedicated task focuses on advancing sentiment analysis techniques for Arabic dialects, specifically in the hotel domain. In fact, Arabic dialects differ significantly in syntax, lexicon, phonology, and semantics, posing serious challenges to NLP. This variability is further compounded in sentiment analysis, where emotional expressions and idiomatic phrases vary widely across regions, making it difficult to achieve consistent sentiment detection. The Ahasis shared task aims to address key challenges in dialect-specific sentiment detection, cross-dialect sentiment consistency, and the nuanced classification of sentiment in Arabic hotel customer reviews.

[1] https://ahasis-42267.web.app/

4 Methodology and Results

We developed a hybrid sentiment analysis model that integrates AraBERT with a custom-built Arabic sentiment lexicon tailored specifically for this task. Given the limited size of the available dataset, we also applied data augmentation to enrich the training data and enhance model robustness. The following subsections describe the development pipeline and modeling steps in detail.

4.1 Dataset Preparation and Preprocessing

Our work is based on the dataset provided by the Ahasis Shared Task, which comprises 860 hotel reviews written in two Arabic dialects: Saudi Arabic and Darija (Moroccan Arabic), with 430 reviews per dialect. Each review is annotated with a sentiment label: positive, neutral, or negative.

4.1.1 Data Augmentation

To enhance model generalization, we applied two complementary data augmentation strategies to the training set. First, a lightweight, custom augmentation method was applied to 20% of the data. This subset was randomly selected to inject controlled lexical variation through word deletion, swapping, and noise injection, ensuring minimal distortion of sentence structure. The selected ratio was chosen to avoid introducing excessive noise while still diversifying the input space.

Second, we used paraphrasing-based augmentation on 50% of the data, employing the AraT5 [2] model (Bani-Almarjeh and Kurdy, 2023) to generate semantically equivalent rephrasings. This larger

[2] almarjeh/t5-arabic-text-summarization

proportion was selected because paraphrased sentences preserve meaning more reliably and therefore can be scaled more safely. As a Transformer-based sequence-to-sequence model designed for Arabic, AraT5 was used to generate fluent, semantically equivalent paraphrases of existing sentences. We fine-tuned the generation parameters, including beam search width, repetition penalties, and output length constraints, to ensure high-quality paraphrase generation.

Both subsets were selected to maintain the original class distribution and ensure no overlap between the two augmented portions. This balanced setup allowed us to maximize training diversity while preserving data quality.

4.1.2 Data Pre-processing

To normalize Arabic input for sentiment classification, we used the ArabertPreprocessor (Wadhawan, 2021), which replicates the preprocessing steps applied during pretraining of AraBERT models. This tool performs a series of operations, including diacritic (Tashkeel) and elongation (Tatweel) removal, normalization of character variants (e.g., forms of Alef), and replacement of URLs, mentions, and emails with special tokens. It also standardizes spacing, and removes redundant characters and punctuation. For dialectal inputs, it helps reduce vocabulary sparsity by enforcing consistent tokenization across variants. These preprocessing steps significantly improve model robustness, especially in low-resource and dialect-rich domains.

4.1.3 Label Encoding and Splitting

Sentiment labels are encoded into numerical representations, by converting categorical sentiment classes (positive, negative, neutral) into a format suitable for model training and evaluation. To maintain representativeness across the training and validation sets, a stratified k-fold cross-validation approach is implemented. This ensures that the proportion of each sentiment class remains consistent across partitions, which is critical for balanced model training.

4.2 Lexicon Building

To enhance the precision of our sentiment classification model, we constructed an Arabic sentiment analysis lexicon using the preprocessed training dataset. Each text sample was thoroughly tokenized and systematically filtered to remove Arabic stopwords. The stopwords list was extended to cover dialect-specific terms relevant to this study, including those from Saudi Arabic and Darija.

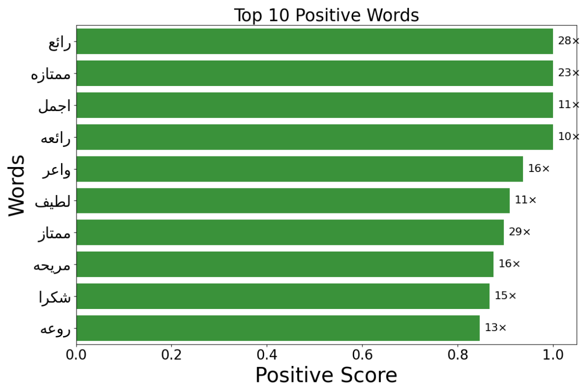


Figure 1: Top 10 Positive Words by Frequency in the Arabic Sentiment Lexicon

Additional non-informative tokens were also removed during this process. Given that the dataset primarily consists of hotel reviews, words such as ”hotel”, ”room”, and ”restaurant” appeared frequently. However, since these terms are context-specific and do not contribute meaningfully to sentiment expression, they were excluded from the final lexicon. The resulting tokens were evaluated based on their frequency within sentiment classes. To ensure statistical significance, tokens with fewer than 10 occurrences were excluded. For each retained token, sentiment scores were calculated based on its relative frequency of occurrence across sentiment-labeled categories in the training dataset. For each word that meets a predefined minimum frequency threshold, we calculate the proportion of times it appears in positive, neutral, and negative samples. We experimented with thresholds of 5, 10, and 15, and found that 10 yielded the best performance. These raw proportions are then normalized so that their

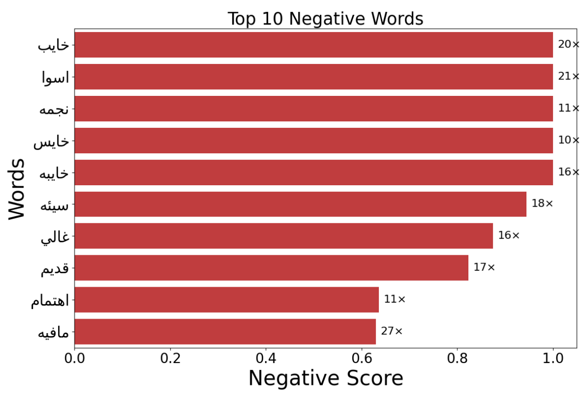


Figure 2: Top 10 Negative Words by Frequency in the Arabic Sentiment Lexicon

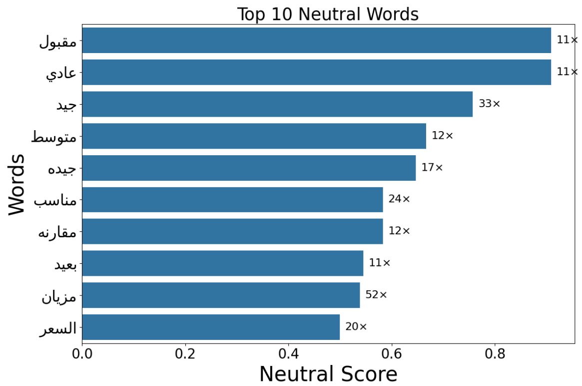


Figure 3: Top 10 Neutral Words by Frequency in the Arabic Sentiment Lexicon

sum equals one, ensuring a probabilistic interpretation of the sentiment distribution. For example, if a word occurs in 18 positive reviews and once each in neutral and negative reviews. This results in raw scores of 0.9 for the positive category (i.e. positive score = 18 / 20 = 0.9) and 0.05 for both neutral and negative (i.e. negative score = 1 / 20 = 0.05), classifying this word as positive in the lexicon since the positive score is the highest.

The finalized lexicon was exported to a CSV file for use in sentiment classification. Each entry in the lexicon includes a token annotated with its corresponding positive, negative, and neutral scores. Visual analyses confirmed the lexicon's coverage and highlighted the most sentiment-representative terms, demonstrating its practical utility. Figures 1, 2, and 3 present the top 20 Arabic words most strongly associated with the positive, negative, and neutral sentiment classes, respectively. These words are selected based on their normalized sentiment scores and frequency in the labeled training dataset. Words are ranked by their sentiment scores. The actual number of occurrences (e.g., 24×) is shown next to each bar. This visualization highlights the most frequent and sentiment-representative terms in each context. This lexicon substantially enhances Arabic sentiment analysis and was successfully integrated into our proposed classifier.

4.3 Lexicon-Based Features Extraction

A comprehensive set of lexical features is extracted from the normalized text, including: positive and negative word counts and ratios, cumulative sentiment scores across the text, statistical measures of sentiment distribution, and presence of sentiment-specific markers identified through linguistic analysis. A sentiment score is then computed for each text instance based on the extracted lexical features. This score serves as both a standalone indicator of sentiment and as an additional feature for the hybrid model, providing interpretable insights into the sentiment orientation of the text.

4.4 Hybrid Model Development

To leverage both contextual embeddings and explicit lexical information, we developed a hybrid neural architecture combining the pre-trained transformer-based language model AraBERT with lexicon-derived features. It consists of a fully-connected layer that transforms the raw lexical features into a dense, lower-dimensional representation that captures the essential sentiment information contained within these features. The contextual embeddings from AraBERT and the processed lexicon features are combined through a concatenation operation, creating a unified representation that leverages both the deep semantic understanding of the transformer model and the explicit sentiment knowledge encoded in the lexicon. The combined features are fed into a classification head consisting of multiple fully-connected layers with non-linear activations and dropout regularization. This component makes the final sentiment classification decision based on the rich, multifaceted representation created through the fusion of contextual and lexical features.

4.5 Results

When evaluated on the test set, our proposed hybrid model achieved an F1-score of 0.74, representing a substantial improvement over the baseline AraBERT-only model, which attained an F1-score of 0.56 (Alharbi et al., 2025b). In addition, the hybrid model showed improved performance in handling sentimentally ambiguous and dialectally diverse hotel reviews. By combining contextual embeddings with lexicon-derived sentiment scores, the model was able to better interpret inputs where polarity cues were subtle or conflicting. Unlike approaches that treat lexical information externally, our method integrates sentiment features directly into the model architecture. Each word's score—calculated from its normalized frequency across sentiment classes—was encoded as a dense feature and fused with AraBERT embeddings. This integration allowed the model to leverage both deep semantic representations and explicit sentiment signals. The lexicon was constructed from the training

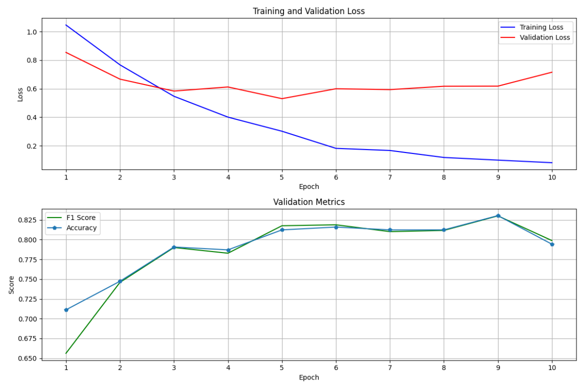


Figure 4: Training and validation performance over 10 epochs for our hybrid model

data, filtered by a minimum frequency threshold, and tailored to Saudi Arabic and Darija, ensuring domain and dialectal relevance. Combined with our data augmentation strategies, this hybrid architecture contributed to greater robustness and improved generalization across dialects. As shown in Figure 4, the top plot shows the training and validation loss curves. The bottom plot on the figure illustrates the upward trend of validation F1-score and accuracy over successive epochs, reflecting improved generalization capabilities of the hybrid model.

5 Conclusion

This paper presents our methodology for Arabic sentiment classification in the context of the multi-dialect hotel review shared task. To address the linguistic complexity and variability of Arabic dialects, we applied comprehensive text normalization using the AraBERT Preprocessor, ensuring consistent representation across dialectal variations. To enhance the training data, we employed two complementary augmentation strategies: (1) a custom probabilistic augmenter that introduced lexical variation through random deletion, token swapping, and controlled noise injection, and (2) a paraphrasing-based approach using the AraT5 model to generate semantically diverse sentence variants. Additionally, we constructed a domain-specific sentiment lexicon that incorporates dialectal vocabulary relevant to the hospitality domain, which was integrated into our hybrid model architecture. The foundation of the hybrid model is built upon AraBERT, which provides deep contextual representations of the input text. In parallel, a dedicated neural network pathway processes the lexicon-derived features through a fully connected layer. The outputs from both pathways are concatenated to form a unified representation, capturing both semantic context and explicit sentiment signals. Our combined approach achieved an F1-score of 74%.

Acknowledgments

We acknowledge Qatar National Research Fund grant NPRP14C0916-210015 from the Qatar Research Development and Innovation Council (QRDI) for funding this research

References

Samia F Abdhood, Nazlia Omar, and Sabrina Tiun. 2025. Data augmentation for arabic text classification: a review of current methods, challenges and prospective directions. *PeerJ Computer Science*, 11:e2685.

Mohamed Elhag Mohamed Abo, Atika Qazi, Ahmed Adel Ahmed Saad, Hager Ali Elsahib, and Ahmed Abdelaziz. 2024. Sentiment analysis in saudi arabic dialect for hajj season services using twitter data. In *2024 1st International Conference on Logistics (ICL)*, pages 1–5.

Souha Al Katat, Chamseddine Zaki, Hussein Hazimeh, Ibrahim El Bitar, Rafael Angarita, and Lionel Trojman. 2024. Natural language processing for arabic sentiment analysis: A systematic literature review. *IEEE Transactions on Big Data*, 10(5):576–594.

Noora Al-shameri and Hend Al-Khalifa. 2024. Arabic paraphrased parallel synthetic dataset. *Data in Brief*, 57:111004.

Amani A Aladeemy, Ali Alzahrani, Mohammad H Algarni, Saleh Nagi Alsubari, Theyazn HH Aldhyani, Sachin N Deshmukh, Osamah Ibrahim Khalaf, Wing-Keung Wong, and Sameer Aqburi. 2024. Advancements and challenges in arabic sentiment analysis: A decade of methodologies, applications, and resource development. *Heliyon*.

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)*.

Abdullah M. Alkadri, Abeer Elkorany, and Cherry Ahmed. 2022. Enhancing detection of arabic social spam using data augmentation and machine learning. *Applied Sciences*, 12(22).

Wael Alosaimi, Hager Saleh, Ali A. Hamzah, Nora El-Rashidy, Abdullah Alharb, Ahmed Elaraby, and Sherif Mostafa. 2024. Arabbert-lstm: improving arabic sentiment analysis based on transformer model and long short-term memory. *Frontiers in Artificial Intelligence*, Volume 7 - 2024.

Mohammad Bani-Almarjeh and Mohamad-Bassam Kurdy. 2023. Arabic abstractive text summarization using rnn-based and transformer-based architectures. *Information Processing Management*, 60(2):103227.

Rehab Duwairi and Ftoon Abushaqra. 2021. Syntactic- and morphology-based text augmentation framework for arabic sentiment analysis. *PeerJ Computer Science*, 7:e469.

Shaista Firdous and Muhammad Saeed Iqbal. 2025. Exploring contemporary arabic sentiment analysis: Methods, challenges, and future trends. *Pakistan Journal of Multidisciplinary Innovation*, 4(1):34–48.

Dhaou Ghoul, Jérémy Patrix, Gaël Lejeune, and Jérôme Verny. 2024. A combined arabert and voting ensemble classifier model for arabic sentiment analysis. *Natural Language Processing Journal*, 8:100100.

Sari Masri, Yaqeen Raddad, Fidaa Khandaqji, Huthaifa I. Ashqar, and Mohammed Elhenawy. 2025. Transformer models in education: Summarizing science textbooks with arabart, mt5, arat5, and mbart. In *Intelligent Systems, Blockchain, and Communication Technologies*, pages 286–300, Cham. Springer Nature Switzerland.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. Arat5: Text-to-text transformers for arabic language generation.

Ruba Obiedat, Duha Al-Darras, Esra Alzaghoul, and Osama Harfoushi. 2021. Arabic aspect-based sentiment analysis: A systematic literature review. *IEEE Access*, 9:152628–152645.

Anshul Wadhawan. 2021. AraBERT and farasa segmentation based approach for sarcasm and sentiment detection in Arabic tweets. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 395–400, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

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 فيهاافتتاح 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.

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

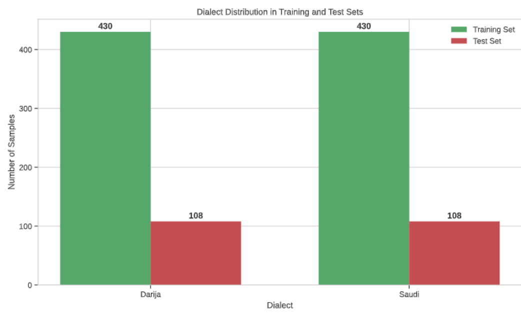


Figure 2: Distribution of dialects in training and test sets

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.

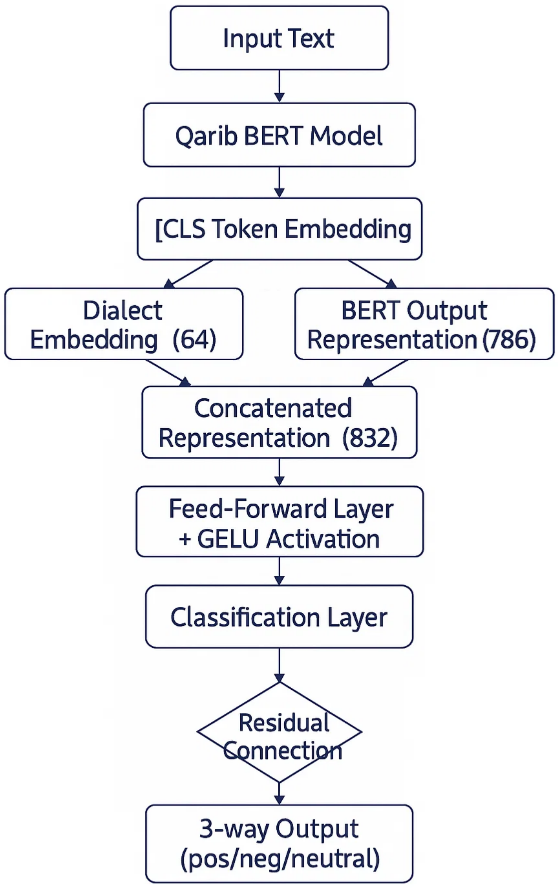


Figure 3: Dialect-Aware Sentiment Analysis Model Architecture

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:

$$\mathrm{GELU}(x) = x \cdot \Phi(x) \tag{1}$$

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

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

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

This approximation provides a smooth, non-linear 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:

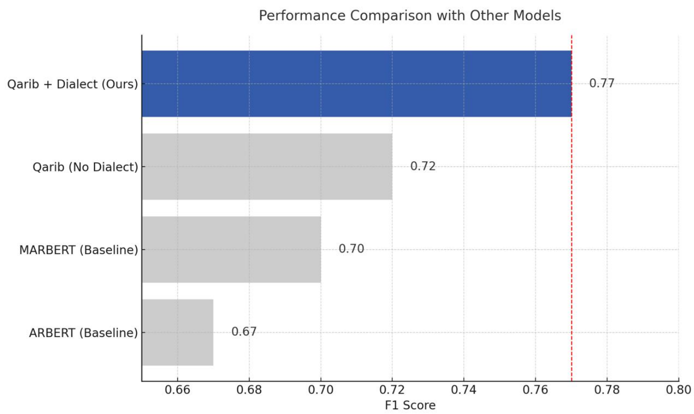


Figure 4: Performance comparison with other models

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.

MAPROC at AHaSIS Shared Task: Few-Shot and Sentence Transformer for Sentiment Analysis of Arabic Hotel Reviews

Randa Zarnoufi
Mohammed V University in Rabat
randa_zarnoufi@um5.ac.ma

Abstract

Sentiment analysis of Arabic dialects presents significant challenges due to linguistic diversity and the scarcity of annotated data. This paper describes our approach to the AHaSIS shared task, which focuses on sentiment analysis on Arabic dialects in the hospitality domain. The dataset comprises hotel reviews written in Moroccan and Saudi dialects, and the objective is to classify the reviewers' sentiment as positive, negative, or neutral. We employed the SetFit (Sentence Transformer Fine-tuning) framework, a data-efficient few-shot learning technique. On the official evaluation set, our system achieved an F1 of 73%, ranking 12th among 26 participants. This work highlights the potential of few-shot learning to address data scarcity in processing nuanced dialectal Arabic text within specialized domains like hotel reviews.

1 Introduction

Sentiment Analysis (SA), a key area within Natural Language Processing (NLP), focuses on identifying and extracting subjective information, such as opinions, emotions, and attitudes, from textual data. The proliferation of user-generated content on social media, review platforms, and forums has underscored the importance of robust sentiment analysis systems across various domains. One such domain is the hospitality industry, where customer reviews significantly influence consumer decisions and provide valuable feedback for service improvement. However, performing sentiment analysis on Arabic text, particularly regional dialects, presents unique and substantial challenges.

Arabic is a morphologically rich language characterized by a complex linguistic landscape. It encompasses Modern Standard Arabic (MSA), used in formal communication and written media, and a diverse array of regional dialects spoken in daily interactions. These dialects often differ significantly from MSA and from each other in terms of syntax, lexicon, phonology, and semantics. This linguistic diversity poses a considerable hurdle for NLP tasks, as models trained on MSA or one specific dialect may not generalize well to others. In the context of sentiment analysis, this variability is further compounded because emotional expressions, idiomatic phrases, and cultural nuances can vary widely across different Arabic-speaking regions, making consistent sentiment detection a difficult endeavor.

The AHaSIS (Arabic Hotel Reviews Analysis for Sentiment Identification in Dialects) shared task (Alharbi, Chafik, et al., 2025) aims to address these challenges by focusing on advancing sentiment analysis techniques specifically for Arabic dialects within the hotel review domain. The primary objective of this task is to classify the overall sentiment expressed in hotel reviews, written in various Arabic dialects, into three categories: positive, neutral, or negative. This task encourages participants to develop models capable of handling the intricacies of dialectal Arabic and accurately discerning sentiment despite the linguistic variations.

This dataset aims to serve as a multidialectal benchmark for sentiment analysis in the hospitality sector, helping to address the shortage of sentiment analysis resources in dialectal Arabic. While previous efforts have contributed valuable

Proceedings of the Shared Task on Sentiment Analysis for Arabic Dialects (AHaSIS) associated with RANLP 2025, pages 46–53, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-109-7-008

benchmarks for SA Arabic benchmarks; such as LABR (Aly & Atiya, 2013) for book reviews, ASTD (Nabil et al., 2015) for Arabic tweets SA, ArSentD-LEV (Baly et al., 2019) for the Levantine dialect on Twitter and Arsen-20 (Fang & Xu, 2024) that focuses on the theme of COVID-19; we still face a pressing need for more diverse and domain-specific datasets.

This paper describes our approach to the AHaSIS shared task. We present a solution that leverages the Huggingface SetFit[1] (Sentence Transformer Fine-tuning) framework, an efficient method for few-shot learning without requiring extensive prompt engineering and faster to train and run inference with, compared to other few-shot learning methods.

The core contribution of our work lies in demonstrating the effectiveness of SetFit for rapidly developing a competitive sentiment analysis model for low-resource dialectal Arabic, achieving promising results with limited training data.

We detail our data preprocessing steps, model configuration, training strategy, and the experimental results obtained on the shared task's dataset in the following sections.

2 Related Work

Sentiment analysis in Arabic has been an active area of research, driven by the increasing volume of Arabic content online and the need for tools to understand public opinion and customer feedback. Early approaches often relied on lexicon-based methods, which utilize dictionaries of words tagged with sentiment polarities such as ArsenL (Arabic Sentiment Lexicon) (Badaro et al., 2014) and Arabic Senti-Lexicon (Al-Moslmi et al., 2018). While straightforward, these methods struggle with the nuances of language, context dependent sentiment, and the morphological richness of Arabic (Mulki et al., 2017). Machine learning techniques, including traditional methods like Support Vector Machines (SVM) (Duwairi et al., 2015), Naive Bayes (Al-Horaibi & Khan, 2016), and Logistic Regression (Alshammari & AlMansour, 2020)have been widely applied to Arabic sentiment analysis. These approaches typically require significant feature engineering, such as n-grams, TF-IDF (Al-Osaimi & Badruddin, 2014; Salameh et al., 2015) and word embeddings. The development of Arabic-specific word embeddings, like AraVec (Soliman et al., 2017), has improved the performance of these models by capturing semantic relationships between words (Ashi et al., 2019). After the introduction of deep learning models, such as CNN and those based on RNNs like LSTMs and GRUs have advanced the SA performance (Al-Sallab et al., 2017; Al-Smadi et al., 2017; Dahou et al., 2016)

More recently, Transformer-based architectures such as AraBERT (Antoun et al., 2020), CAMeLBERT (Inoue et al., 2021), and ARBERT/MARBERT (Abdul-Mageed et al., 2021), which are specifically trained on large Arabic and dialectal corpora, have achieved state-of-the-art results in various NLP tasks, including SA. Pre-training language models for Arabic and fine-tuning it for SA, significantly improves the performance (Eljundi et al., 2019). These models can learn contextual representations of text, reducing the need for manual feature engineering.

Despite these advancements, SA for Arabic dialects remains a significant challenge (Shi & Agrawal, 2025). Most existing resources and models are primarily focused on Modern Standard Arabic (MSA) or a limited set of well-resourced dialects (Mashaabi et al., 2024). The linguistic diversity across dialects, including variations in vocabulary, grammar, and idiomatic expressions, makes it difficult to develop universally applicable sentiment analysis tools. Furthermore, the scarcity of large, annotated datasets for many Arabic dialects hinders the development and evaluation of robust dialectal sentiment analysis models.

Few-shot learning techniques have emerged as a promising direction for addressing data scarcity in NLP. Methods like SetFit that stands for Sentence Transformer Finetuning (Tunstall et al., 2022) which is employed in our work, aim to achieve strong performance with minimal labeled training examples. SetFit leverages sentence transformers to generate high-quality embeddings and then fine-tunes a classification head, offering an efficient alternative to training large models from scratch or relying on complex prompting strategies. SetFit uses supervised contrastive learning that has proved its efficiency in previous works (Khosla et al., 2020).

Our work builds upon these advancements by applying SetFit to the specific challenge of

[1] https://huggingface.co/docs/setfit/en/index

Dialect	Sentence	Polarity
Darija	عطاوني حتى كيكة فعيد ميلادي منين كنت جالس تماك. يستاهل كل فلس حطيتيه They even gave me a cake on my birthday while I was sitting there. It deserves every penny you spent	positive
	الأتمنة دلوطيل طالعة بزاف، وما كاين حتى مقابل فالسيرفيس ولا باش يرضيو الكليان The hotel prices are way too high, and there's nothing in the service to match them or to please the customers	negative
	فطور ما بيهش، يمكن يكون احسن The breakfast wasn't bad, but it could be better	neutral
Saudi	غرف الفندق اطلالتها ساحرة ، نظافة الغرف مع دورات المياه وكل شي في الفندق حلو مره The hotel rooms have a stunning view, the rooms and bathrooms are clean, and everything about the hotel is just great	positive
	غالي مره ملعب اطفال صغير فريق ترفيه خايس Very expensive, the kids' play area is small, and the entertainment team is terrible	negative
	فندق لازمه اهتمام ممكن يكون جيد بس يحتاج تعديلات كثيرة The hotel needs attention, it could be good, but it requires a lot of improvements.	neutral

Table 1: Data Examples from the Training Set.

Dataset	Size	Sentiment Distribution
Training	860 samples: 430 samples for each dialect (Darija/Saudi)	Positive: 308 = 35.81% Negative: 336 = 39.06% Neutral: 216 = 25.11%
Evaluation	216 samples: 108 for each dialect	—

Table 2: Overall Data Distribution

Dialect	Sentiment Distribution
Darija/Saudi	Positive: 154 = 35.81% Negative: 168 = 39.06% Neutral: 108 = 25.11%

Table 3: Sentiment Distribution over Dialects in the Training Data.

sentiment analysis in diverse Arabic dialects within the hotel review domain, contributing to the growing body of research on low-resource NLP and dialectal Arabic processing.

3 Data

The primary dataset for this work was provided as part of the AHaSIS shared task. The dataset consists of hotel reviews written in various Arabic dialects, namely Moroccan dialect known as Darija and Saudi dialect. Some reviews are shown in Table 1 for both dialects.

The training set is composed of 860 samples, and the evaluation set contains 216 samples. The training set contains four columns for 'ID' of review, 'Sentiment' (the sentiment label: positive, neutral, or negative), 'Text' (the review text), and 'Dialect' (Darija or Saudi). The overall data distribution and sentiment class distribution over dialects are described in Table 2 and Table 3 respectively. We can observe that data distribution and the sentiment distribution over dialects are perfectly balanced 50% for each one. Although the sentiment classes (positive, negative, neutral) are not fully balanced, the differences in their distribution are relatively small and unlikely to introduce significant bias into the overall classification.

4 System Description

Our approach to the AHaSIS shared task on sentiment analysis for Arabic dialects in hotel reviews is centered around the SetFit (Sentence Transformer Fine-tuning) framework. SetFit is designed for efficient few-shot learning, enabling

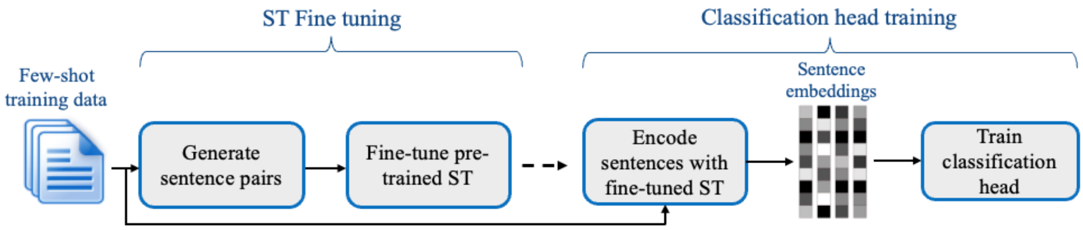


Figure 1: The SetFit framework consists of two processing stages: the first involves fine-tuning a sentence transformer, and the second trains a classification head on the resulting embeddings.

the development of robust text classification models with a small number of labeled examples, which are particularly beneficial for low-resource languages or specialized domains like dialectal Arabic.

4.1 SetFit Framework Overview

The SetFit framework is composed of two main parts: the body which is a sentence transformer model, and a classification head as it is shown in Figure 1[2]. It operates in two stages. First, it fine-tunes the pre-trained sentence transformer model using a contrastive learning objective on the provided training data. This stage aims to generate rich sentence embeddings directly from text examples that are well-suited for the specific task and dataset. In this process, positive pairs (similar sentences) and negative pairs (dissimilar sentences) are created from the training examples to teach the model to produce similar embeddings for sentences with the same label and dissimilar embeddings for sentences with different labels. The second stage involves training a classification head (typically a logistic regression model or a simple neural network) on top of the frozen embeddings generated by the fine-tuned Sentence Transformer. This two-stage process allows SetFit to achieve strong performance without requiring extensive labeled data or complex prompt engineering, which are often associated with large language models.

4.2 Sentence Transformer Model

For our implementation, we utilized a pre-trained Arabic sentence transformer model Arabic-SBERT-100K[3]. This model is a variant of SBERT (Sentence BERT), that it is based on AraBERT base model and has been finetuned on a large corpus of Arabic text (100K sentences), making it suitable for generating meaningful sentence representations for Arabic. The choice of this model was based on its reported performance in comparison with other models and its availability on the Hugging Face Model Hub. For comparison purposes, we also used another sentence transformer model which is ArabicBERT_Finetuned-AR-500[4]

4.3 Baseline Model

We initiated our baseline by fine-tuning AraBERTv0.2 base model[5] on the full training dataset and then explored ways to enhance classification performance by varying the number of training samples.

4.4 Dataset

For our experiments, we used the provided train.csv file. While analyzing the data we have observed that it has been already pre-processed and doesn't need further pre-processing except removing punctuation and normalizing some Arabic letters like alif and hamza writing format (‘ا’, ‘أ’, ‘إ’).

We performed an 80/20 split on this dataset to create our internal training and testing partitions. For the few-shot learning aspect of SetFit, a smaller training subset was created by sampling a specific number of examples per class from the training set.

Our final reported model utilized 64 examples per sentiment class for fine-tuning the sentence transformer and training the classification head. This means the actual training data used by SetFit was relatively small.

[2] https://huggingface.co/blog/setfit
[3] https://huggingface.co/akhooli/Arabic-SBERT-100K
[4] https://huggingface.co/danfeg/ArabicBERT_Finetuned-AR-500
[5] https://huggingface.co/aubmindlab/bert-base-arabertv2

Model	Samples Number	Epoch	F1 %	Duration in h:m:s
Baseline: finetuned AraBERT	Full training dataset	3	75.07	00:17:10
Arabic-SBERT-100K	8	1	63.82	00:05:12
	8	3	67.48	00:15:14
	8	5	64.22	00:26:09
	16	3	69.19	00:30:54
	32	3	73.84	01:04:35
	64	3	**78.87**	02:16:12
ArabicBERT_Finetuned-AR-500	8	3	52.83	00:43:32

Table 4: Results on the AHaSIS Training dataset.

Model	Samples Number	Epoch	F1
Arabic-SBERT-100K	32	3	72%
	64	3	**73%**

Table 5: Results on the AHaSIS Evaluation dataset.

4.5 Training Procedure

The training process followed the standard SetFit pipeline. We selected a subset of the training data for the few-shot learning setup. We experimented with different numbers of training examples per class.

The SetFitTrainer was configured with the following key parameters:

Model: The Arabic-SBERT-100K SetFit model.

Training Dataset: The few-shot training dataset containing ***n*** examples per class.

Test Dataset: The full test dataset derived from the initial data split.

Number of Epochs: The sentence transformer was fine-tuned for 3 epochs during the contrastive learning phase. This value was chosen based on preliminary experiments, which showed that 3 epochs provided a good balance between performance and training time for this specific model and dataset size.

Batch Size: A batch size of 16 was used during the fine-tuning of the sentence transformer.

Number of Iterations: 20 iterations were used for generating text pairs for contrastive learning.

Loss Function: CosineSimilarityLoss was employed for the contrastive fine-tuning stage, which is a standard choice for SetFit to encourage similar sentences to have embeddings with high cosine similarity.

After the Sentence Transformer was fine-tuned, the classification head was trained on the embeddings generated from the selected training set samples.

For the final reported model, we used 64 examples per class from the training set to fine-tune the sentence transformer and train the classification head. This corresponds to the shot train set created by sampling 64 instances for each label ('positive', 'negative', 'neutral').

5 Results and Discussion

This section details the results achieved by our SetFit-based sentiment analysis model for Arabic hotel reviews. We mentioned that for our experimental settings, all experiments were conducted using CPU (i7-11850H) with 8 cores and 32GB of RAM to prove that this approach is not computationally expensive. The timing of each experience is also reported in Table 4.

5.1 Results

The primary metric for the shared task is F1 score.

The key results relevant to our submitted approach are shown in Table 5 and are as follows:

Using Arabic-SBERT-100K with 64 samples per class and 3 epochs for fine-tuning we achieved a F1 score of 73% in the official evaluation set, significantly outperforming the shared task baseline of 56% (Alharbi, Ezzini, et al., 2025).

We report also the specific results achieved with different configurations based on our internal test set split in Table 4 and for each dialect and class in Table 6.

These results demonstrate a clear trend:

- Increasing the number of few-shot samples per class, up to 64 in these experiments, generally improves the model's

Dialect	Class	F1 %
Darija	positive	72.92
	negative	**85.85**
	neutral	61.30
Saudi	positive	**87.52**
	negative	85.55
	neutral	80.00

Table 6: Results per Dialect and Class on the AHaSIS Training dataset.

performance on the internal test set. The choice of 3 epochs for fine-tuning the sentence transformer also appears to be beneficial compared to a single epoch and five epochs when using a small number of samples.

- The results obtained using the SetFit framework with the Arabic-SBERT-100K sentence transformer model were better than those with ArabicBERT_Finetuned-AR-50.
- Our internal experiments, culminating in an F1 of 78.87% on our held-out test set when using 64 samples per class, suggest that SetFit can effectively leverage pre-trained Arabic language representations for this task with relatively minimal training data.

One of the key strengths of our approach is its efficiency. Traditional deep learning models often require substantial amounts of labeled data and extensive computational resources for training. SetFit, by fine-tuning a sentence transformer with a contrastive objective and then training a simple classification head, offers a more resource-friendly alternative.

The ability to achieve competitive performance with only 64 examples per class (a total of 192 training examples for three classes) underscores the data efficiency of this method.

We mentioned that the official leaderboard on the shared task website[6] and Codabench show various teams' F1 scores, providing a benchmark for performance.

5.2 Discussion

The progressive improvement in F1 as the number of few-shot samples per class increased (from 8 to 64 samples) aligns with expectations. More data, even in a few-shot context, generally allows the model to learn more robust representations and decision boundaries. The choice of 3 epochs for fine-tuning the sentence transformer also appeared to be beneficial, suggesting that even a brief period of contrastive learning can adapt the pre-trained embeddings effectively to the target task and domain.

Comparing our F1 score of 73% with the best F1 scores on the official AHaSIS leaderboard (where the top score is 81% F1), our model appears to be competitive. Several factors could influence the performance. The quality of the pre-trained sentence transformer is crucial. Its training on a substantial Arabic corpus likely provides a good foundation for understanding Arabic semantics and syntax. The diversity of dialects present in the AHaSIS dataset is a significant challenge. While SetFit helps, the model's ability to generalize across highly varied dialectal expressions might still be limited by the relatively small fine-tuning dataset. The hotel review domain also has its specific vocabulary and expressions of sentiment, which the model needs to learn.

As we mentioned, we had also evaluated the model on our internal test set for each dialect and class. The results in Table 6 show that the model performs best on the Saudi dialect, achieving over 80% F1 in all classes, including strong handling of neutral sentiment (80%). In contrast, it struggles more with Darija, especially in the neutral class (61.3%), suggesting difficulty in capturing subtler expressions in that dialect which may be underrepresented in the sentence transformer training data.

In our experiments we also tested an alternative model, ArabicBERT_Finetuned-AR-500, which achieved a lower F1 (52.83 % with 3 epochs). This highlights the importance of selecting an appropriate base sentence transformer model for the SetFit framework. Arabic-SBERT-100K model seems better suited for this specific task based on these preliminary results.

[6] https://Ahasis-42267.web.app/leaderboard

Finally, A detailed error analysis on the predictions would be valuable to understand the types of reviews or dialects where the model performs poorly, guiding further improvements.

6 Conclusion

In this paper, we presented our solution for the AHaSIS shared task, which focused on sentiment analysis of Arabic hotel reviews written in various dialects. Our approach utilized the SetFit framework, a few-shot learning technique. This methodology was chosen for its efficiency in scenarios with limited labeled data, a common challenge in processing diverse Arabic dialects.

Our system achieved an F1 of 73% in the official evaluation set. This performance suggests that few-shot learning with appropriate pre-trained models is a viable strategy for tackling sentiment analysis in complex linguistic landscapes like dialectal Arabic. The results also indicated a positive correlation between the number of few-shot samples and model performance, within the tested range.

Future work could explore several avenues for improvement. Experimenting with a wider range of pre-trained Arabic or multilingual sentence transformers, conducting a more extensive hyperparameter optimization for the SetFit trainer, and increasing the number of few-shot training examples could potentially enhance performance.

References

Abdul-Mageed, M., Elmadany, A., Moatez, E., & Nagoudi, B. (2021). *ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic*. https://github.com/attardi/wikiextractor.

Alharbi, M., Chafik, S., Ezzini, S., Mitkov, R., Ranasinghe, T., & Hettiarachchi, H. (2025). AHaSIS: Shared Task on Sentiment Analysis for Arabic Dialects. *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Alharbi, M., Ezzini, S., Hettiarachchi, H., Ranasinghe, T., & Mitkov, R. (2025). Evaluating Large Language Models on Arabic Dialect Sentiment Analysis. *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Al-Horaibi, L., & Khan, M. B. (2016). Sentiment analysis of Arabic tweets using text mining techniques. *International Workshop on Pattern Recognition*. https://api.semanticscholar.org/CorpusID:59094364

Al-Moslmi, T., Albared, M., Al-Shabi, A., Omar, N., & Abdullah, S. (2018). Arabic senti-lexicon: Constructing publicly available language resources for Arabic sentiment analysis. *Journal of Information Science*, *44*(3), 345–362. https://doi.org/10.1177/0165551516683908

Al-Osaimi, S., & Badruddin, K. M. (2014). Role of Emotion icons in Sentiment classification of Arabic Tweets. *Proceedings of the 6th International Conference on Management of Emergent Digital EcoSystems*, 167–171. https://doi.org/10.1145/2668260.2668281

Al-Sallab, A., Baly, R., Hajj, H., Shaban, K. B., El-Hajj, W., & Badaro, G. (2017). AROMA: A Recursive Deep Learning Model for Opinion Mining in Arabic as a Low Resource Language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, *16*(4). https://doi.org/10.1145/3086575

Alshammari, N. F., & AlMansour, A. A. (2020). *Aspect-based Sentiment Analysis for Arabic Content in Social Media*.

Al-Smadi, M., Qawasmeh, O., Al-Ayyoub, M., Jararweh, Y., & Gupta, B. B. (2017). Deep Recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews. *J. Comput. Sci.*, *27*, 386–393. https://api.semanticscholar.org/CorpusID:51918752

Aly, M., & Atiya, A. (2013). LABR: A Large Scale Arabic Book Reviews Dataset. In H. Schuetze, P. Fung, & M. Poesio (Eds.), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 494–498). Association for Computational Linguistics. https://aclanthology.org/P13-2088/

Antoun, W., Baly, F., & Hajj, H. (2020). *AraBERT: Transformer-based Model for Arabic Language Understanding*. http://arxiv.org/abs/2003.00104

Ashi, M. M., Siddiqui, M. A., & Nadeem, F. (2019). *Pre-trained Word Embeddings for Arabic Aspect-Based Sentiment Analysis of Airline Tweets* (A. E. Hassanien, M. F. Tolba, K. Shaalan, & A. T. Azar, Eds.; Vol. 845). Springer International Publishing. https://doi.org/10.1007/978-3-319-99010-1

Badaro, G., Baly, R., Hajj, H., Habash, N., & El-Hajj, W. (2014). A Large Scale Arabic Sentiment Lexicon for Arabic Opinion Mining. In N. Habash & S. Vogel (Eds.), *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)* (pp. 165–173). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3623

Baly, R., Khaddaj, A., Hajj, H. M., El-Hajj, W., & Shaban, K. B. (2019). ArSentD-LEV: A Multi-Topic Corpus for Target-based Sentiment Analysis in Arabic Levantine Tweets. *CoRR*, *abs/1906.01830*. http://arxiv.org/abs/1906.01830

Dahou, A., Xiong, S., Zhou, J., Haddoud, M. H., & Duan, P. (2016). Word Embeddings and Convolutional Neural Network for Arabic Sentiment Classification. In Y. Matsumoto & R. Prasad (Eds.), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 2418–2427). The COLING 2016 Organizing Committee. https://aclanthology.org/C16-1228/

Duwairi, R., Ahmed, N. A., & Al-Rifai, S. Y. (2015). Detecting sentiment embedded in Arabic social media - A lexicon-based approach. *J. Intell. Fuzzy Syst.*, *29*, 107–117. https://api.semanticscholar.org/CorpusID:6145765

Eljundi, O., Antoun, W., Nour,), Droubi, E., Hajj, H., El-Hajj, W., & Shaban, K. (2019). *hULMonA ( ): The Universal Language Model in Arabic*.

Fang, Y., & Xu, C. (2024). ArSen-20: A New Benchmark for Arabic Sentiment Detection. *5th Workshop on African Natural Language Processing*. https://openreview.net/forum?id=GgsRUF5kJt

Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., & Habash, N. (2021). *The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models*. http://arxiv.org/abs/2103.06678

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., & Krishnan, D. (2020). *Supervised Contrastive Learning*. http://arxiv.org/abs/2004.11362

Mashaabi, M. O., Al-Khalifa, S. Z., & Al-Khalifa, H. S. (2024). *A Survey of Large Language Models for Arabic Language and its Dialects*. https://mistral.ai/news/mistral-saba

Mulki, H., Haddad, * -Hatem, & Babao˘, I. (2017). *Modern Trends in Arabic Sentiment Analysis: A Survey* (Vol. 58).

Nabil, M., Aly, M., & Atiya, A. (2015). ASTD: Arabic Sentiment Tweets Dataset. In L. Màrquez, C. Callison-Burch, & J. Su (Eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 2515–2519). Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1299

Salameh, M., Mohammad, S. M., & Kiritchenko, S. (2015). *Sentiment after Translation: A Case-Study on Arabic Social Media Posts*. http://www.purl.com/net/ArabicSentiment

Shi, Z., & Agrawal, R. (2025). *A comprehensive survey of contemporary Arabic sentiment analysis: Methods, Challenges, and Future Directions*.

Soliman, A. B., Eissa, K., & El-Beltagy, S. R. (2017). AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP. *Procedia Computer Science*, *117*, 256–265. https://doi.org/10.1016/j.procs.2017.10.117

Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., & Pereg, O. (2022). *Efficient Few-Shot Learning Without Prompts*. http://arxiv.org/abs/2209.11055

mucAI at Ahasis Shared Task: Sentiment Analysis with Adaptive Few Shot Prompting

Ahmed Abdou
Technical University of Munich
ahmed.abdou@dtum.de

Abstract

Sentiment Analysis is a crucial task in Natural Language Processing (NLP) focused on identifying and categorizing emotional tones or opinions within text. For Arabic customer reviews, sentiment analysis is particularly challenging. The language's rich diversity, with numerous regional dialects differing significantly from Modern Standard Arabic (MSA) and each other in lexicon, syntax, and sentiment expression, complicates consistent performance across dialects. In this paper, we present our approach, submitted to the AHASIS Shared Task 2025, focusing on sentiment analysis for Arabic dialects in the hotel domain. Our method leverages the capabilities of GPT-4o through adaptive few-shot prompting technique, where similar contextual examples are dynamically selected for each review using a k-Nearest Neighbors (kNN) search over train embeddings from a fine-tuned encoder model. This approach tailors the prompt to each specific instance, enhancing classification performance over minority class. Our submission achieved an F1-score of 76.0% on the official test set, showing stronger performance for the Saudi dialect compared to Darija.

1 Introduction

Sentiment analysis for the Arabic language presents unique challenges due to its complex linguistic landscape. Unlike languages with more homogeneous structures, Arabic encompasses Modern Standard Arabic (MSA) and numerous regional dialects that differ in syntax, lexicon, morphology, and semantic expressions. These variations become particularly pronounced when analyzing sentiment in domain-specific contexts, such as hotel reviews, where emotional expressions and idiomatic phrases can vary significantly across dialectal boundaries. In this paper, we tackle the AHASIS shared task on sentiment analysis on arabic dialects (Saudi and Darija) in the hospitality domain.

Large Language Models (LLMs) exhibit impressive in-context learning (ICL) abilities: with a handful of demonstrations in the prompt they adapt to new tasks on the fly (Brown et al., 2020). Yet a growing body of work shows that ICL is highly sensitive to which and how many examples are shown (Yoshida, 2024). Small, static prompts amplify demonstration bias: models over-predict labels that dominate the prompt or appear later in the example list, harming minority classes. Traditional approaches have often relied on fine-tuning pre-trained transformer models specific to Arabic, such as AraBERT (Antoun et al., 2020) and MARBERT (Abdul-Mageed et al., 2021), which have shown considerable success. However, the advent of LLMs like has opened new frontiers, offering powerful generative and reasoning capabilities. For example, using LLMs we can get the generated tokens by the model justifying the given classification label (Huang et al., 2023).

Motivated by these findings, we explore the use of GPT-4o (Hurst et al., 2024) for the AHASIS shared task. We evaluate three prompting strategies: (i) zero-shot prompting, where no examples are provided; (ii) static few-shot prompting, using a fixed set of manually curated demonstrations; and (iii) adaptive few-shot prompting, where examples are dynamically retrieved from the training set based on semantic similarity using AraBERT embeddings. Importantly, each retrieved example is paired with a GPT-4o-generated chain of thought conditioned on its gold label. For comparison, we also fine-tune AraBERT and MARBERT as encoder-based baselines. Our key contribution is the adaptive few-shot strategy, which consistently outperformed both static and zero-shot prompting on the test set. By tailoring examples to each input, this method offers a simple yet effective way to improve LLM performance in multilingual and

Proceedings of the Shared Task on Sentiment Analysis for Arabic Dialects (AHaSIS) associated with RANLP 2025, pages 54–61, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-109-7-009

dialectal sentiment tasks.

2 Related Work

LLMs have recently approached or matched the performance of supervised task models without gradient updates. A systematic evaluation (Zhang et al., 2024) across 26 datasets showed GPT-3.5/4 and Llama-2 (Touvron et al., 2023) within 1–3 macro-F1 of fine-tuned RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2020) baselines on sentence- and aspect-level sentiment. Building on this, several studies have evaluated general-purpose LLMs on Arabic sentiment tasks. Taqyim (Alyafeai et al., 2023) benchmarked GPT-3.5 and GPT-4 across seven Arabic datasets and found that GPT-4 narrows but does not eliminate the gap to supervised MARBERT (Abdul-Mageed et al., 2021). GPTAraEval (Khondaker et al., 2023) evaluated ChatGPT [1] on 44 Arabic tasks and reported consistent under-performance relative to smaller, Arabic-tuned pre-trained language models (PLMs). Moreover, in (Al-Thubaity et al., 2023), the authors showed that GPT-4 in a 5-shot setting reaches AraBERT (Antoun et al., 2020) performance, whereas GPT-3.5 and Google Bard (PaLM 2) (Anil et al., 2023) lag behind. A comparative evaluation (Alharbi et al., 2025b) of DeepSeek-R1 (Guo et al., 2025), Qwen2.5 (Qwen et al., 2025), and LLaMA-3 (Grattafiori et al., 2024) further demonstrates the efficacy of dialect-specific prompting and parameter-efficient fine-tuning (LoRA) (Hu et al., 2022) in Arabic sentiment analysis, showing that prompt-input alignment significantly enhances performance, especially for underrepresented dialects.

kNN-Prompting (Xu et al., 2023) embeds all training instances once and predicts each test query by a simple majority vote over its k closest neighbours, thus bypassing context-length limits and heavy calibration steps. kNN-ICL (Zhao et al., 2023) refines this idea by selecting similarity-based demonstrations on the fly, yielding consistent F1 gains under strict token budgets. The closest antecedent to our pipeline is OpenMedLM (Maharjan et al., 2024), which retrieves the five most similar patient questions, asks an LLM to write gold-label-conditioned rationales once, caches them, and inserts the triple (question, chain-of-thoughts (CoT), answer) at inference—achieving state-of-the-art accuracy on medical QA without fine-tuning.

[1] https://openai.com/index/chatgpt/

3 Task and Data

The AHASIS shared task (Alharbi et al., 2025a) involves sentiment classification of hotel reviews, specifically targeting Arabic dialects—Saudi and Darija. Each review in the dataset is annotated with two labels: the dialect (Saudi or Darija) and the sentiment expressed (positive, negative, or neutral). The primary objective of the task is to predict the sentiment label given a sentence in one of these dialects.

The provided training dataset consists of 860 sentences, evenly split with 430 sentences for each dialect. Within each dialect subset, there are 154 positive, 168 negative, and 108 neutral sentiment sentences. To conduct our experiments, we stratified the training data based on both sentiment and dialect, splitting it into 80% training and 20% development subsets obtaining 688 sentences for training and 172 sentences for development. Additionally, a separate test set comprising 216 sentences was provided, with annotations specifying only the dialect and excluding sentiment labels.

4 Method

Building on insights from prior work on retrieval-augmented prompting and chain-of-thought (CoT) reasoning, we introduce a dynamic *k*NN + label-conditioned CoT framework. Unlike static prompting methods, which rely on fixed demonstrations and a global decision boundary, our approach constructs an adaptive, query-specific prompt at inference time.

For each input review, we (i) embed it using a pre-trained Arabic encoder, (ii) retrieve its k nearest dialect-balanced neighbours from a cached set of labelled hotel reviews, (iii) generate a rationale for each neighbour using GPT-4o, conditioned on its gold label, and (iv) assemble these (review, label, CoT) triplets into the prompt.

We evaluate our approach alongside two encoder-based baselines (AraBERT and MARBERT), and compare three prompting strategies using GPT-4o: zero-shot, static few-shot, and our proposed adaptive few-shot prompting. All models are tested on a shared train/dev/test split to ensure comparability across settings.

4.1 Encoder-Based Baselines

We fine-tune two established Arabic transformer models, **AraBERT** and **MARBERT**, as classification baselines. These serve as strong nongenerative benchmarks for comparison with GPT-4o-based approaches.

4.2 Zero-Shot and Static Few-Shot Prompting

In the zero-shot setup, GPT-4o receives only a system instruction in Arabic asking it to: (1) assign one of the three sentiment labels (positive, neutral, negative), and (2) generate a brief natural language justification for its prediction.

For the static few-shot setting, we prepend a fixed set of curated examples—each containing an input review, its gold label, and a justification—to the prompt. This improves performance but has limited flexibility: when test instances deviate aspect-wise from the static examples, performance degrades. Additionally, expanding the static prompt to cover more cases is costly and difficult to maintain.

4.3 Adaptive Few-Shot Prompting

To address the limitations of static prompting, we propose an **adaptive few-shot strategy** that builds a tailored prompt for each input review. The process consists of two stages:

(1) Retrieval: For a given input review x, we compute its embedding using a pre-trained AraBERT encoder and retrieve the $k = 20$ nearest neighbors from the training set. These are then stratified by sentiment label, and we select the top $n = 3$ examples per class (if available), yielding up to 9 demonstrations. This approach balances semantic similarity and aspect alignment (e.g., topic or focus of the review) with label diversity.

(2) Prompt Construction with CoT: Each retrieved example is paired with a chain-of-thought (CoT) justification generated by GPT-4o, conditioned on its gold label (see prompt in Table 8). These structured examples—(review, label, justification)—are inserted into the prompt, followed by the target review x to be classified. This dynamic prompt ensures that each input is evaluated in a context shaped by semantically and topic relevant reasoning chains.

5 Results

In this section, we present the experimental results obtained across different prompting strategies using GPT-4o, alongside baseline results from the encoder models AraBERT and MARBERT. Our primary metric for evaluation is the macro-F1 score. Table 1 summarizes the overall results on both the development and test sets. The code used for the experiments is available on GitHub [2].

Model / Method	Dev	Test
AraBERT	83.7	73
MARBERT	85.2	73
Zero-Shot	80.5	75
Few-Shot	82.5	74
Adaptive Few-Shot	**84.5**	**76**

Table 1: Macro-F1 scores for sentiment classification across different experimental setups.

Table 2: Class level F1-score for Saudi dialect sentences

Prompting Strategy	Neg	Neu	Pos	Macro
Zero-Shot	96	70	84	83
Few-Shot	97	78	88	88
Adaptive Few-Shot	**97**	**81**	**89**	**89**

Table 3: Class level F1-score for Darija dialect sentences

Prompting Strategy	Neg	Neu	Pos	Macro
Zero-Shot	88	63	82	78
Few-Shot	89	62	83	78
Adaptive Few-Shot	**90**	**67**	**84**	**80**

Our adaptive few-shot method achieved the highest test set performance among the GPT-4o prompting techniques, with an F1-score of 76%. This placed our method sixth in the AHASIS shared task leaderboard. Interestingly, although MARBERT achieved the highest development set score of 85.2%, its performance dropped noticeably to 73% on the test set, similar to AraBERT.

Tables 2 and 3 provides a detailed breakdown of F1-scores for each sentiment class (Negative, Neutral, Positive) and the overall Macro F1-score, comparing our three prompting strategies across the Saudi and Darija dialects on the development set. Consistent with the overall test set performance noted in Table 1, the Adaptive Few-Shot strategy yielded the highest Macro F1-scores for both dialects: 89 for Saudi and 80 for Darija.

A key objective of the adaptive strategy was to address challenges with the neutral class. For the Saudi dialect, the Neutral F1-score improved substantially from 70 (Zero-shot) and 78 (Few-shot) to 81 with Adaptive Few-Shot. Similarly, for the Darija dialect, Adaptive Few-Shot improved the Neutral F1-score to 67 compared to (Zero-shot) 63

[2] https://github.com/AhmedAbdel-Aal/Ahasis_shared_task

and (Few-Shot) 62. Performance on the Negative and Positive classes was generally strong across all methods, particularly for the Saudi dialect where Zero-shot already achieved F1-scores of 96 (Negative) and 84 (Positive). The Adaptive Few-Shot method largely maintained or slightly enhanced these high scores while making its most significant impact on the Neutral class. Interestingly, the impact of the standard Few-Shot prompting varied by dialect. While it offered clear improvements for the Saudi dialect, increasing the Macro F1 from 83 (Zero-shot) to 88, it provided no overall benefit for the Darija dialect, where the Macro F1 remained at 78, and the Neutral F1-score even saw a slight decrease.

Comparing the two dialects, the overall F1-scores for Darija were consistently slightly lower than those for Saudi across all prompting methods. For instance, with Adaptive Few-Shot, the Macro F1 was 89 for Saudi versus 80 for Darija.

6 Discussion

The superior performance of the adaptive few-shot strategy can be attributed to its dynamic, instance-specific contextualization. Traditional fine-tuning creates a static, global decision boundary that applies uniformly across all test inputs, which can be suboptimal for linguistically diverse or ambiguous cases. Similarly, static few-shot prompting relies on a fixed set of demonstrations that may not align well with the semantics of a given test instance, limiting their ability to guide the model effectively. In contrast, our adaptive few-shot approach constructs query-specific prompt for each input. This adaptation enables the model to better capture subtle distinctions—particularly near the boundary between neutral and positive sentiment—by grounding its reasoning in semantically relevant examples. In effect, improving precision and recall for neutral class, see Table 4.

Table 4: Precision (P) and Recall (R) per sentiment class (%) across prompting strategies.

Strategy	Negative		Neutral		Positive	
	P	R	P	R	P	R
Zero-Shot	86	99	78	58	83	84
Few-Shot	89	97	81	60	82	89
Adaptive Few-Shot	89	99	85	65	85	89

More precisely, under zero-shot prompting, the neutral class achieved a precision of 78%, indicating that when the model predicted neutral, it was often correct. However, the low recall of 58% resulted in a modest F1-score of 67%. This suggests that the model relied on a general understanding of what counts as "neutral" in language, rather than learning how neutrality is defined in this specific dataset. As a result, it often misclassified factual or mildly opinionated reviews as positive (see Table 6 for an example).

Introducing static few-shot prompting led to moderate improvements. For the neutral class, precision improved slightly to 81%, and recall increased marginally to 60%, resulting in an F1-score of 69%. The modest gain in recall suggests that manually selected examples help the model better recognize prototypical neutral instances, but may still fall short in capturing the full diversity of this class. The most significant gains came from adaptive few-shot prompting, where performance improved across all classes. The neutral class in particular benefited, with precision rising to 85% and recall improving to 65%, leading to its highest F1-score of 74%. This shows that providing contextually relevant examples helped the model better handle ambiguity and make more consistent decisions. Compared to static prompts, the adaptive strategy offered examples that were closer in meaning and tone to the input, which helped the model better understand what neutrality looks like in this dataset. Reviews that contained mild opinions or balanced descriptions—previously misclassified as positive—were more often labeled correctly. This suggests that dynamic prompting helped the model adjust its decision boundary more accurately around the neutral class (see Table 7 for an example).

Given the computational overhead introduced by adaptive few-shot prompting, we analyze its runtime and token-level cost in practice. The final classification prompt includes 3 to 9 demonstrations per test instance, averaging 778 input and 239 output tokens. This range results from selecting up to 3 examples per sentiment class (negative, neutral, positive) from the 20 nearest neighbors; if the retrieved set lacks class diversity, fewer than 9 examples are included. These demonstrations are generated using justification generation prompt shown in Table 5, with an average of 95 input and 219 output tokens each. We cache all intermediate generations, so each training example is used for CoT generation at most once. In total, each test instance, on average, incurs one main LLM call (778 in + 239 out) and 3 to 9 smaller calls (95 in + 219 out each).

7 Limitations and Future Work

The effectiveness of the adaptive few-shot strategy is closely tied to the capabilities of the underlying language model (GPT-4o) and the quality of the encoder used for kNN retrieval. Limitations in the LLM's understanding of specific dialects, or in the encoder's ability to generate semantically meaningful embeddings, may propagate through the retrieval process and affect final predictions. In this work, the encoder choice was not extensively optimized, leaving room for improvement in retrieval quality. Additionally, we used semantic similarity as a proxy for selecting aspect-aligned reviews, but this approximation may not always capture the most relevant examples for each input—especially when sentiment is conveyed through subtle tone, emphasis, or implied preferences rather than explicit aspect terms. Our experiments were conducted exclusively with GPT-4o, and it remains unclear how well the observed improvements from adaptive prompting would transfer to other large language models. In addition, we did not perform extensive hyperparameter tuning or architecture exploration for the encoder-based baselines (AraBERT and MARBERT). These models were fine-tuned with standard settings to provide a comparative reference, but stronger results might be achievable with more targeted optimization. Finally, key components of our system—such as the number of neighbors (k), the number of few-shot examples, and the structure of the prompts—were selected based on preliminary experiments rather than exhaustive tuning. A more systematic exploration of hyperparameters and prompt formats could further enhance performance and provide insight into robustness and generalization.

An important future direction is to evaluate the impact of chain-of-thought generation by comparing our approach to a version of adaptive few-shot prompting that uses retrieved examples without CoT. This would help isolate the contribution of the reasoning component and better understand its role in guiding sentiment classification. Moreover, our experiments were limited to GPT-4o; testing this approach with open-source Arabic LLMs would provide insight into its generalizability and practicality in low-resource or non-proprietary settings.

8 Conclusion

In this shared task, we tackled sentiment analysis for Arabic dialects within the hotel domain, focusing on the AHASIS 2025 dataset. Our approach centered on inspecting the capabilities of the LLMs, specifically GPT-4o, through various prompting strategies. We introduced an adaptive few-shot prompting technique, where during inference, we dynamically selected relevant contextual examples for each review by performing a kNN search over cached embeddings from the training set, which were generated by a fine-tuned Arabert model. This approach aimed to improve generalization and address challenges like the neutral class by providing similar aspects, relevant context for each specific instance. The results demonstrate a clear improvement with the incorporation of our adaptive few-shot prompting. Specifically, on the test set, the Macro-F1 score achieved was 76.0% with the adaptive approach, surpassing the 75.0% from zero-shot and 74.0% from static few-shot prompting, as well as our fine-tuned encoder baselines which scored 73.0%. Our submission secured 6th place in the shared task leaderboard.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for arabic. In *Proc. ACL*, pages 7088–7105.

Abdulmohsen Al-Thubaity, Sakhar Alkhereyf, Hanan Murayshid, Nouf Alshalawi, Maha Omirah, Raghad Alateeq, Rawabi Almutairi, Razan Alsuwailem, Manal Alhassoun, and Imaan Alkhanen. 2023. Evaluating ChatGPT and bard AI on Arabic sentiment analysis. In *Proceedings of ArabicNLP 2023*, pages 335–349, Singapore (Hybrid). Association for 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)*.

Zaid Alyafeai, Maged S. Alshaibani, Badr AlKhamissi, Hamzah Luqman, Ebrahim Alareqi, and Ali Fadel.

2023. Taqyim: Evaluating arabic nlp tasks using chatgpt models.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *Proc. OSACT*, pages 9–15.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H Gilpin. 2023. Can large language models explain themselves? a study of llm-generated self-explanations. *arXiv preprint arXiv:2310.11207*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Gptaraeval: A comprehensive evaluation of chatgpt on arabic nlp. *arXiv preprint arXiv:2305.14976*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jenish Maharjan, Anurag Garikipati, Navan Preet Singh, Leo Cyrus, Mayank Sharma, Madalina Ciobanu, Gina Barnes, Rahul Thapa, Qingqing Mao, and Ritankar Das. 2024. Openmedlm: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. *Scientific Reports*, 14(1):14156.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Benfeng Xu, Quan Wang, Zhendong Mao, Yajuan Lyu, Qiaoqiao She, and Yongdong Zhang. 2023. $k$nn prompting: Beyond-context learning with calibration-free nearest-neighbour inference. In *Proc. ICLR*.

Lui Yoshida. 2024. The impact of example selection in few-shot prompting on automated essay scoring using gpt models. In *International Conference on Artificial Intelligence in Education*, pages 61–73. Springer.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

Wenting Zhao, Ye Liu, Yao Wan, Yibo Wang, Qingyang Wu, Zhongfen Deng, Jiangshu Du, Shuaiqi Liu, Yunlong Xu, and Philip S Yu. 2023. Knn-icl: Compositional task-oriented parsing generalization with nearest neighbor in-context learning. *arXiv preprint arXiv:2312.10771*.

Appendix: Prompts and Model Outputs

A.1 Justification Generation Prompt

Table 5: Prompt used to generate chain-of-thought (CoT) justifications for each retrieved review in the adaptive few-shot setting.

Justification Generation Prompt

```
Question: Justify the sentiment of the following
Saudi dialect hotel review.
Review: {review}
Label: {label}
Return your response in the following XML format:
<analysis>
 <overall_justification> your step-by-step
analysis </overall_justification>
</analysis>
```

A.2 Zero-Shot and Adaptive Few-Shot Prompt Examples

We show the GPT-4o outputs for the same development review under two prompting strategies: zero-shot and adaptive few-shot. The overall prompt structure remains consistent across all settings; the only difference lies in the addition of review examples in the few-shot and adaptive versions. In this case, the zero-shot prompt leads to an incorrect prediction, while the adaptive few-shot prompt produces the correct label. This illustrates how retrieved examples—selected based on both semantic similarity and aspect alignment—help the model better calibrate its decision boundary, particularly in distinguishing between neutral and positive sentiment.

Table 6: Example of zero-shot prompting for sentiment classification in the Saudi dialect.

Prompt

Question: what is the sentiment of the following hotel review in Saudi dialect?

Review: رخيص وفي احسن موقع حصلت خصم لاني اشتغل في جهة تابعة القطاع الخاص و لكن حتى الاسعار العادية بدون الخصم رخيصة بالنسبة للموقع
Return your response in the following XML format:
<analysis>
<overall_sentiment> positive/neutral/negative </overall_sentiment>
<overall_justification> your step-by-step analysis </overall_justification>
</analysis>

GPT-4o Output

<analysis>
<overall_sentiment> positive </overall_sentiment>
<overall_justification>
The review highlights several positive aspects of the hotel. Firstly, the reviewer mentions that the hotel is رخيص (cheap), which is generally a positive attribute for budget-conscious travelers. Secondly, the location is described as في احسن موقع (in the best location), indicating a favorable opinion. The reviewer also notes receiving a discount, and even mentions that regular prices are cheap for the location. Overall, the review expresses satisfaction with pricing and location.
</overall_justification>
</analysis>

True Label: Neutral
Predicted Label (Zero-Shot): Positive (Incorrect)

Table 7: Example of adaptive few-shot prompting for sentiment classification in the Saudi dialect. For brevity, only one retrieved example is shown.

Prompt (Excerpt)

Question: what is the sentiment of the following hotel review in Saudi dialect?
Examples:

Review: جودة بالنسبة للسعر بالنسبة لمدينة الاسكندرية من افضل الفنادق بهذا السعر و السعر مناسب
<analysis>
<overall_sentiment> Neutral </overall_sentiment>
<overall_justification>

1. The review states جودة بالنسبة للسعر which translates to "Quality for the price," indicating a positive aspect of the hotel. 2. The phrase افضل الفنادق بهذا السعر translates to "one of the best hotels at this price," which further emphasizes a positive sentiment regarding the hotel's value. 3. The review mentions السعر مناسب meaning "the price is appropriate," reinforcing the idea that the reviewer finds the pricing reasonable. 4. However, the overall tone of the review is factual and lacks strong emotional language or expressions of enthusiasm, which is often associated with positive sentiment. 5. The use of neutral language and the focus on price and quality without expressing strong feelings leads to the conclusion that the sentiment is neutral rather than overtly positive.
</overall_justification>
</analysis>
Now, analyze the following review:

Review: رخيص وفي احسن موقع حصلت خصم لاني اشتغل في جهة تابعة القطاع الخاص و لكن حتى الاسعار العادية بدون الخصم رخيصة بالنسبة للموقع
Return your response in the following XML format...

GPT-4o Output

<analysis>
<overall_sentiment> neutral </overall_sentiment>
<overall_justification>

1. The review begins with آرخيص وفي احسن موقع, indicating positive language.
2. Mentions receiving a discount for private sector work, reinforcing affordability.
3. Even without the discount, prices are considered cheap for the location.
4. However, the tone is factual and lacks strong emotion.
5. Overall, the review is descriptive rather than opinionated, suggesting neutrality.

</overall_justification>
</analysis>

True Label: Neutral
Predicted Label (Adaptive Few-Shot): Neutral (Correct)

A Hybrid Transformer-Based Model for Sentiment Analysis of Arabic Dialect Hotel Reviews

Rawand Alfugaha
College of Information Technology
Lusail University
Doha, Qatar
ralfoqha@lu.edu.qa

Mohammad AL-Smadi
Digital Learning and Online Education Office
Qatar University
Doha, Qatar
malsmadi@qu.edu.qa

Abstract

This paper describes the AraNLP system developed for the "Ahasis" shared task on sentiment detection in Arabic dialects for hotel reviews. The task involved classifying the overall sentiment of hotel reviews (Positive, Negative, or Neutral) written in Arabic dialects, specifically Saudi and Darija. Our proposed model, AraNLP, is a hybrid deep learning classifier that leverages the strengths of a transformer-based Arabic model (AraELECTRA) augmented with classical bag-of-words style features (TF-IDF). Our system achieved an F1-score of 76%, securing the 5th rank in the shared task, significantly outperforming the baseline system's F1-score of 56%.

1 Introduction

Arabic dialect sentiment analysis presents unique challenges due to morphological complexity, diglossia, and regional variations (Abdul-Mageed et al., 2021). While Modern Standard Arabic (MSA) has been well-studied, dialects like Saudi and Darija remain under-resourced despite their prevalence in user-generated content (Salameh et al., 2018; Talafha et al., 2020). Recent advances in transformer models have shown promise for Arabic NLP (Antoun et al., 2020), but dialect-specific adaptations remain limited. Hotel reviews are particularly challenging due to domain-specific terminology mixed with dialectal variations (AL-Smadi et al., 2023).

The Ahasis shared task (Alharbi et al., 2025a) presented a significant challenge in the field of Arabic Natural Language Processing (NLP), focusing on sentiment analysis in the hospitality domain, specifically for diverse Arabic dialects. Sentiment analysis of user-generated content, such as hotel reviews, provides invaluable insights for both businesses and consumers (Al-Smadi et al., 2019). However, the Arabic language, with its rich morphology and wide range of dialects, poses unique difficulties for NLP tasks. Modern Standard Arabic (MSA) is the formal version of the language, while numerous regional dialects (e.g., Egyptian, Levantine, Gulf, Maghrebi) are predominantly used in informal online communication, including hotel reviews. These dialects often lack standardized orthography and can differ significantly from MSA and from each other in terms of lexicon, syntax, and morphology (Birjali et al., 2021).

In this paper, we present our system, AraNLP, which participated in the "Ahasis" shared task. Our approach is a hybrid deep learning model that combines the contextual understanding capabilities of a pre-trained transformer-based model for Arabic (AraELECTRA) with the statistical strength of TF-IDF features. This hybrid architecture aims to capture both semantic nuances and important lexical cues from the review texts.

The rest of this paper is organized as follows: Section 2 sheds the light on related work, Section 3 demonstrates the research methodology, Section 4 presents the model results, Section 5 discusses the model results, and Section 6 concludes the research paper and provides insights for future work.

2 Related Work

Sentiment analysis in Arabic, particularly for dialectal Arabic, has garnered increasing attention from the research community. Early approaches often relied on lexicon-based methods, which utilize predefined dictionaries of words tagged with sentiment polarities (Birjali et al., 2021). While straightforward, these methods struggle with the nuances of dialects, context-dependent sentiment, and the lack of comprehensive dialectal lexicons.

Machine learning techniques, including Support Vector Machines (SVM), Naive Bayes, and Logis-

Proceedings of the Shared Task on Sentiment Analysis for Arabic Dialects (AHaSIS) associated with RANLP 2025, pages 62–68, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-109-7-010

tic Regression, have been widely applied to Arabic sentiment analysis, often outperforming lexicon-based approaches when sufficient labeled data is available (Al-Smadi et al., 2018). These models typically rely on features such as n-grams, TF-IDF, and word embeddings. For instance, Al-Smadi et al. (2019) explored the use of morphological, syntactic, and semantic features to enhance aspect-based sentiment analysis of Arabic hotel reviews, demonstrating the value of linguistic features.

Deep learning models, particularly those based on Recurrent Neural Networks (RNNs) like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs), have shown significant promise in capturing sequential information and contextual dependencies in text (Alyami et al., 2022). Elfaik and Nfaoui (2020) and Ombabi et al. (2020) employed LSTM networks for aspect-based sentiment analysis of Arabic text, highlighting their effectiveness. More recently, AL-Smadi et al. (2023) proposed a GRU model combined with a multilingual universal sentence encoder for Arabic aspect-based sentiment analysis, achieving strong results.

Transformer-based models, such as BERT and its variants, have revolutionized the field of NLP by achieving state-of-the-art performance on various tasks, including sentiment analysis. Several pre-trained transformer models have been developed specifically for the Arabic language, such as ALBERT (Lan et al., 2019), AraBERT (Antoun et al., 2020), AraELECTRA (Antoun et al., 2021), QARiB(QCRI Arabic and dialectal BERT) (Abdelali et al., 2021), and CAMeLBERT (Inoue et al., 2021). These models are pre-trained on large Arabic corpora and can be fine-tuned for specific downstream tasks like sentiment classification. The use of such models is becoming increasingly common due to their ability to understand complex linguistic patterns and contextual information. Recent work has also explored hybrid approaches combining transformers with other neural network architectures. For example, Bourahouat et al. (2024) proposed BERT-based models that are pre-trained on Arabic datasets, namely AraBERT, QARIB, ALBERT, AraELECTRA, and CAMeLBERT integrated with machine learning and deep learning models such as SVM and CNN for sentiment analysis of Darija (Moroccan dialect).

Hybrid models combining transformers with sequential or ensemble components have gained traction. (Alzahrani et al., 2024) achieved 97% accuracy by integrating AraBERT with LSTM to model long-term dependencies in Arabic text. For dialect detection, (Saleh et al., 2025) proposed a stacked transformer framework (AraBERT and XLM-R) with a meta-learner, achieving 93% F1-score on the IADD dataset. In sentiment analysis, (Mansour et al., 2025) demonstrated that transformer ensembles outperform single-model approaches by aggregating linguistic features across dialects.

The Ahasis shared task builds upon this body of work by focusing on the challenging aspects of dialectal Arabic (Saudi and Darija) in the specific domain of hotel reviews. Our work contributes to this line of research by proposing a hybrid model that combines the strengths of transformer architectures with traditional feature engineering to tackle the sentiment analysis task in diverse Arabic dialects.

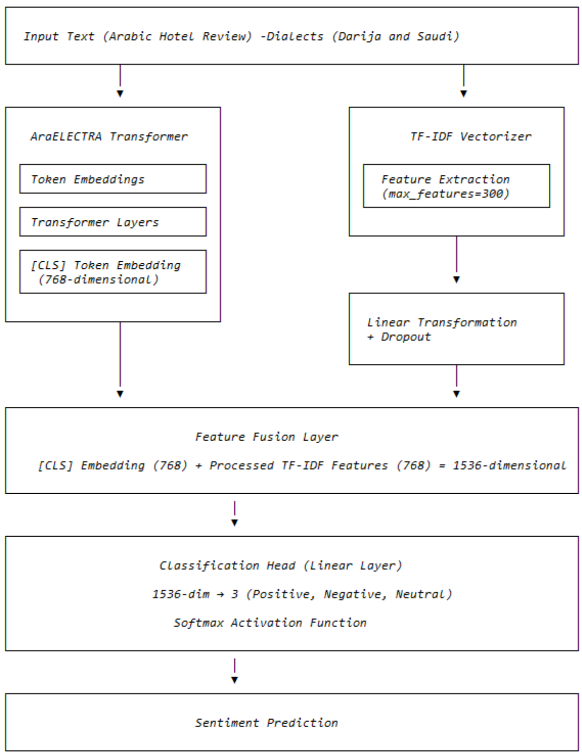


Figure 1: AraNLP Model Architecture: Hybrid integration of AraELECTRA and TF-IDF for sentiment classification of Arabic dialect hotel reviews.

3 Research Methodology

This section outlines the methodology employed in developing the AraNLP system. We first describe the shared task, followed by details of the dataset provided and our model architecture.

3.1 Task Definition

The Ahasis Shared Task focuses on sentiment classification in the hospitality domain for Arabic dialects. Specifically, given a hotel review written in either Moroccan Arabic (Darija) or Saudi dialect, the goal is to predict its overall sentiment as Positive, Neutral, or Negative (Alharbi et al., 2025a). Unlike aspect-based sentiment analysis which targets sentiment toward specific aspects (Alyami et al., 2022; AL-Smadi et al., 2023), this task concerns the general sentiment of the entire review. The official evaluation metric is Macro-averaged F1-score across the three sentiment classes, ensuring that performance on each class (including the often under-represented neutral class) is given equal importance. Participants were provided a labeled dataset (with a predefined train/test split) and a baseline model for reference. The baseline using AraBERT attained 56% Macro-F1, illustrating the difficulty of capturing sentiment in this domain and setting a performance bar for participants (Alharbi et al., 2025b).

3.2 Dataset

The shared task dataset consists of Arabic hotel reviews collected from online sources (e.g., booking websites or social media platforms). The training set contains 860 reviews and the test set contains 216 reviews. Both sets are evenly balanced across the two dialects and the three sentiment categories. In practice, this means the training data has roughly equal numbers of Moroccan Darija and Saudi reviews (approximately 430 each), and within each dialect the distribution of positive, neutral, and negative labels is also approximately equal. The reviews vary in length from short comments (a few words) to longer sentences. Some examples of typical review content include praise or complaints about the room, cleanliness, staff behavior, price, or location. Neutral reviews often describe the experience factually without strong emotion. Positive reviews might use enthusiastic phrases or adjectives (in dialect, e.g., "زوين بالزاف" meaning "very nice" in Darija), whereas negative reviews contain criticism or negative expressions (e.g., "ما عجبنيش الحال" meaning "I did not like the situation" in Darija, or "مو نظيف أبداً" meaning "not clean at all" in Saudi dialect).

3.3 Data Preprocessing

We performed only tokenization. We used an Arabic tokenizer (compatible with AraELECTRA's vocabulary) to segment each review into tokens. We did not apply stemming, lemmatization, dialect normalization, or remove stopwords. The rationale was to let the AraELECTRA model and TF-IDF vector to capture the presence of any word that might carry sentiment (including shifting words, dialect words or foreign terms). While more aggressive text normalization (e.g., unifying Arabic letter variants or removing diacritics) can sometimes help, we chose to keep the text intact to preserve dialectal cues (for instance, the difference between "جميل" (beautiful in MSA) and "زوين" (beautiful in Darija) is important to maintain). The dataset was used in the given train/test split; we did not use cross-validation or external data. A small portion of the training set was held out as a validation set for early stopping and model selection, as described below.

3.4 Model

Our proposed system, AraNLP, employs a hybrid deep learning architecture designed to effectively capture both semantic and lexical features from Arabic hotel reviews. The core components of our model are a pre-trained transformer model (AraELECTRA) and TF-IDF features, which are combined and passed through a classification head.

3.4.1 AraELECTRA Embeddings

We utilize AraELECTRA (Antoun et al., 2021), a transformer-based model pre-trained on a large corpus of Arabic text. AraELECTRA is an ELECTRA-style model, which is trained as a discriminator to distinguish between original input tokens and plausible but synthetically generated replacements produced by a small generator network. This pretraining scheme has been shown to be more sample-efficient than standard masked language modeling (MLM) approaches like BERT. For each input review, we feed the tokenized text into AraELECTRA to obtain contextualized embeddings for each token. We use the embedding of the special '[CLS]' token as the aggregate representation of the review's semantics.

3.4.2 TF-IDF Features

To complement the deep contextual features from AraELECTRA, we incorporate traditional bag-of-words style features using Term Frequency-Inverse

Document Frequency (TF-IDF). We use 'TfidfVectorizer' to convert the collection of review texts into a matrix of TF-IDF features. We set the maximum number of features (i.e., dimensionality of the TF-IDF vectors) to 300. We tried different dimensions (i.e. 100 and 500) but 300 achieved the best results. These features capture the importance of different words in distinguishing between sentiment classes based on their frequency in individual documents and across the entire corpus.

3.4.3 Feature Fusion and Classification

The AraELECTRA '[CLS]' token embedding and the 300-dimensional TF-IDF vector are first processed independently. The TF-IDF vector is passed through a linear transformation layer followed by a dropout layer to project it into a space that is compatible with the transformer embeddings and to add regularization. The resulting processed TF-IDF vector is then concatenated with the AraELECTRA '[CLS]' embedding. This combined feature vector, which now contains both rich semantic information from the transformer and salient lexical information from TF-IDF, is then fed into a final linear classification layer with a softmax activation function to predict the sentiment class (Positive, Negative, or Neutral).

3.4.4 Training Setup

We trained our AraNLP model using the *AdamW optimizer* with a learning rate of 2e-5. A linear warm-up scheduler was employed for the learning rate. The loss function used was *CrossEntropyLoss*, with equal weighting for all three sentiment classes to handle potential class imbalances. We implemented an early stopping mechanism based on the validation loss. We monitored the validation loss after each epoch. If the validation loss did not improve for 3 consecutive epochs, we stopped training. In practice, our model converged within 5 epochs. We found that validation loss typically plateaued or began to increase after the 4th epoch. Early stopping helped prevent overfitting on spurious patterns in the training set. The model parameters from the epoch with the lowest validation loss were retained for final evaluation on the test set.

4 Results

Our AraNLP system was evaluated on the official test set provided by the Ahasis shared task organizers. The primary evaluation metric was the macro F1-score, which considers the F1-score for each sentiment class (Positive, Negative, Neutral) and then averages them, providing a balanced measure of performance across all classes.

As depicted in Table 1, AraNLP, achieved a macro F1-score of 76%. This performance placed our system at the 5th rank among all participating teams in the shared task. For comparison, the baseline system provided by the Ahasis organizers (referred to as "BaseLine (Ahasis)") achieved a macro F1-score of 56% and was ranked 13th. This indicates that our hybrid approach, combining AraELECTRA with TF-IDF features, provided a substantial improvement of 20 percentage points in F1-score over the baseline. The accuracy achieved by our system was also recorded, though the primary ranking was based on the macro F1-score. Detailed per-class precision, recall, and F1-scores, if provided by the organizers or obtainable from our experiment logs, would offer further insights but are summarized here by the macro F1-score.

Table 1 summarizes the key results of our system in comparison to the baseline.

5 Discussion

The performance of our AraNLP system, achieving a macro F1-score of 76% and ranking 5th in the Ahasis shared task, is encouraging. The substantial improvement over the baseline (56% F1) highlights the efficacy of our hybrid approach. The fusion of contextual embeddings from AraELECTRA with traditional TF-IDF features appears to provide a synergistic effect, capturing both deep semantic understanding and salient lexical cues. AraELECTRA, pre-trained on a vast Arabic corpus, offers robust representations of Arabic text, including dialectal variations to some extent. The TF-IDF features, on the other hand, can effectively highlight words that are strongly indicative of a particular sentiment, which might be particularly useful for domain-specific jargon or highly polar expressions not fully captured by the general pre-training of the transformer.

The challenges inherent in Arabic dialect sentiment analysis, such as the lack of standardized orthography, code-switching, and the nuanced expression of sentiment, are significant (Birjali et al., 2021). Our model's ability to perform well despite these challenges suggests that the combination of pre-trained transformers and carefully selected classical features is a promising direction. The 300-dimensional TF-IDF vector, passed through a linear

System	Macro F1-Score (%)	Rank
AraNLP (Our System)	76	5
BaseLine	56	13

Table 1: Performance comparison of AraNLP-SENT with the baseline system on the Ahasis shared task test set.

transformation and dropout, likely helped in regularizing the model and projecting these sparse features into a denser space that could be effectively combined with the transformer embeddings.

However, the error analysis reveals several limitations of our current model. We analyzed a subset of development set instances where AraNLP's prediction was incorrect, to understand the failure modes. The following examples highlight four such misclassifications, along with possible reasons:

Example 1: الفندق جميل ولكن الخدمة سيئة جدا – True label: Negative; Predicted: Positive. This is a code-mixed sentiment within one sentence: "The hotel is beautiful but the service is very bad." Our model likely picked up on the word "جميل" ("beautiful") as a strong positive indicator from both AraELECTRA and TF-IDF perspectives. The presence of "سيئة جدا" ("very bad") should denote negativity, but it appears the model either gave more weight to the positive part or failed to properly model the contrast introduced by "لكن" ("but"). This suggests difficulty in handling sentences with mixed sentiment. A better handling of contrastive conjunctions or a more fine-grained sentiment analysis (aspect-based) might be needed to get these correct.

Example 2: ما عجبنيش الثمن، الغرفة صغيرة بزاف – True label: Negative; Predicted: Neutral. This Moroccan Darija review translates to: "I did not like the price; the room is very small." It is clearly negative, complaining about cost and room size. The model predicted Neutral, possibly because the sentence structure is slightly complex (with negation "ما عجبنيش" meaning "did not please me") and multiple issues listed. AraELECTRA might have struggled with the dialect negation construct ("...ش" suffix) if it wasn't common in pretraining data. Also, "بزاف" ("very") amplifies negativity but without a direct negative word next to it, the model might not strongly connect it to negative sentiment.

Example 3: الغرفة مو بطّالة لكن التكييف مزعج طوال الليل – True label: Positive; Predicted: Negative. This Saudi dialect sentence means: "The room is not bad, but the air conditioning is noisy all night." The model predicted Negative, likely focusing on the complaint. The phrase "مو بطّالة" ("not bad") is faint praise. The model may have been confused by "مزعج" ("noisy"), without understanding that a minor complaint does not negate overall satisfaction.

Example 4: كان المتوقع أفضل بكثير – True label: Negative; Predicted: Neutral. This short review means: "The expected (experience) was much better." It implies disappointment. There is no frankly written negative word; the sentiment is implicit. This suggests a limitation in understanding nuanced or implied sentiment.

These examples illustrate that while AraNLP handles straightforward language well, it can stumble on contrast, negation, and mixed sentiments. Improvements could include contrast modeling, sentiment lexicons, or more training data, especially in dialects. Lastly, sentiment is inherently subjective (See Example 3). Some reviews are borderline. A multi-label or continuous sentiment score model might better reflect these cases in future work. A better accurate solution and better reflecting the value of of customers review is using aspect-based sentiment analysis (Pontiki et al., 2016).

While our system performed well, there is still room for improvement. The gap between our F1-score and those of the top-ranked systems suggests that further refinements could be beneficial. One area for future exploration could be more sophisticated feature fusion techniques. Instead of simple concatenation, attention mechanisms could be employed to allow the model to dynamically weigh the importance of transformer embeddings versus TF-IDF features for different inputs. Additionally, incorporating other linguistic features, such as those derived from morphological analysis of dialectal reviews, might provide further gains, as suggested by prior work like (Al-Smadi et al., 2019).

Another aspect to consider is the handling of neutral reviews. Often, neutral sentiment is harder to classify as it can encompass a wider range of expressions, including factual statements, mixed opinions, or irrelevant content. Analyzing the per-class performance, if available, could shed light on whether our model struggled more with the neutral class compared to positive and negative classes. Tailoring specific strategies for neutral class detection or employing a hierarchical classification approach might be beneficial.

6 Conclusion

In this paper, we presented AraNLP, a hybrid deep learning system for sentiment analysis of Arabic hotel reviews, developed for the ”Ahasis” shared task. Our model combines the strengths of the pre-trained AraELECTRA transformer model with classical TF-IDF features to classify sentiment in diverse Arabic dialects, specifically Saudi and Darija. The AraNLP-SENT system achieved a macro F1-score of 76%, securing the 5th rank and significantly outperforming the baseline system. This result underscores the effectiveness of integrating deep contextual embeddings with traditional lexical features for tackling the complexities of Arabic dialect sentiment analysis.

Future work will focus on exploring more advanced feature fusion techniques, incorporating dialect-specific linguistic resources, and investigating methods to better handle the nuances of neutral sentiment expressions. Further research into larger and more diverse dialectal datasets will also be crucial for advancing the field of Arabic sentiment analysis.

References

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training bert on arabic tweets: Practical considerations. *arXiv preprint arXiv:2102.10684*.

Muhammad Abdul-Mageed, Abdelrahim Elmadany, et al. 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.

Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yaser Jararweh, and Omar Qawasmeh. 2019. Enhancing aspect-based sentiment analysis of arabic hotels' reviews using morphological, syntactic and semantic features. *Information Processing & Management*, 56(2):308–319.

Mohammad AL-Smadi, Mahmoud M. Hammad, Sa'ad A. Al-Zboon, Saja AL-Tawalbeh, and Erik Cambria. 2023. Gated recurrent unit with multilingual universal sentence encoder for arabic aspect-based sentiment analysis. *Knowledge-Based Systems*, 261:107540.

Mohammad Al-Smadi, Omar Qawasmeh, Mahmoud Al-Ayyoub, Yaser Jararweh, and Brij Gupta. 2018. Deep recurrent neural network vs. support vector machine for aspect-based sentiment analysis of arabic hotels' reviews. *Journal of computational science*, 27:386–393.

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)*.

Salha Alyami, Areej Alhothali, and Amani Jamal. 2022. Systematic literature review of arabic aspect-based sentiment analysis. *Journal of King Saud University-Computer and Information Sciences*, 34(9):6524–6551.

Mohammed Alzahrani, Ashraf Elnagar, and James O'Shea. 2024. Arabert-lstm: Improving arabic sentiment analysis based on transformer model and long short-term memory. *Frontiers in Artificial Intelligence*, 7:1408845.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. Araelectra: Pre-training text discriminators for arabic language understanding. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 191–195.

Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107134.

Ghizlane Bourahouat, Manar Abourezq, and Najima Daoudi. 2024. Improvement of moroccan dialect sentiment analysis using arabic bert-based models. *J. Comput. Sci*, 20(2):157–167.

Hanane Elfaik and El Habib Nfaoui. 2020. Deep bidirectional lstm network learning-based sentiment analysis for arabic text. *Journal of Intelligent Systems*, 30(1):395–412.

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. *arXiv preprint arXiv:2103.06678*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Omar Mansour, Eman Aboelela, Remon Talaat, and Mahmoud Bustami. 2025. Transformer-based ensemble model for dialectal arabic sentiment classification. *PeerJ Computer Science*, 11:e2644.

Abubakr H Ombabi, Wael Ouarda, and Adel M Alimi. 2020. Deep learning cnn–lstm framework for arabic sentiment analysis using textual information shared in social networks. *Social Network Analysis and Mining*, 10:1–13.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *International workshop on semantic evaluation*, pages 19–30.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained arabic dialect identification. In *Proceedings of the 27th international conference on computational linguistics*, pages 1332–1344.

Hager Saleh, Abdulaziz AlMohimeed, Rasha Hassan, Mandour M Ibrahim, Saeed Hamood Alsamhi, Moatamad Refaat Hassan, and Sherif Mostafa. 2025. Advancing arabic dialect detection with hybrid stacked transformer models. *Frontiers in Human Neuroscience*.

Bashar Talafha, Mohammad Ali, Muhy Eddin Za'ter, Haitham Seelawi, Ibraheem Tuffaha, Mostafa Samir, Wael Farhan, and Hussein Al-Natsheh. 2020. Multi-dialect arabic bert for country-level dialect identification. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 111–118.

Arabic-Centric Large Language Models for Dialectal Arabic Sentiment Analysis Task

Salwa Alahmari[1,3], Eric Atwell[1], Hadeel Saadany[2] and Mohammed Alsalka[1]

[1]University of Leeds, UK
scssala, E.S.Atwell,m.a.alsalka@leeds.ac.uk
[2]Birmingham City University , UK
hadeel.saadany@surrey.ac.uk
[3]University of Hafr Al Batin, Saudi Arabia
ssalahmari@uhb.edu.sa

Abstract

This paper presents a study on sentiment analysis of Dialectal Arabic (DA), with a particular focus on Saudi and Moroccan (Darija) dialects within the hospitality domain. We introduce a novel dataset comprising 698 Saudi Arabian proverbs annotated with sentiment polarity labels—Positive, Negative, and Neutral—collected from five major Saudi dialect regions: Najdi, Hijazi, Shamali, Janoubi, and Sharqawi. In addition to this, we used customer reviews for fine-tuning the CAMeLBERT-DA-SA model, which achieved a 75% F1 score in sentiment classification. To further evaluate the robustness of Arabic-centric models, we assessed the performance of three open-source large language models—Allam, ACeGPT, and Jais—in a zero-shot setting using the Ahasis shared task test set. Our results highlight the effectiveness of domain-specific fine-tuning in improving sentiment analysis performance and demonstrate the potential of Arabic-centric LLMs in zero-shot scenarios. This work contributes new linguistic resources and empirical insights to support ongoing research in sentiment analysis for Arabic dialects

1 Introduction

Arabic ranks as the fourth most commonly used language on the Internet, spoken by over 400 million individuals(Guellil et al., 2021), and serves as the official language across 22 nations(Farghaly and Shaalan, 2009). Known for its complex and richly structured morphology, Arabic exists in three primary forms: Modern Standard Arabic (MSA), Classical Arabic (CA), and a wide range of regional dialects(Al-Sulaiti and Atwell, 2006; Guellil et al., 2021). Each Arabic-speaking country typically has one or more local dialects, adding layers of complexity for researchers working with the language. Sentiment analysis involves evaluating individuals' opinions and emotions about products, services, organizations, people, and other subjects by analyzing textual data. This process classifies text into positive, negative, or neutral categories to measure public sentiment. Social media platforms serve as a crucial data source for sentiment analysis, given their extensive use for expressing opinions and sharing information. With the continuous rise in social media users, the volume of data available for sentiment analysis is also expanding. Sentiment analysis in dialectal Arabic presents numerous unique challenges due to the language's rich diversity and complexity. Dialectal Arabic varies significantly across regions, reflecting distinct phonological, lexical, and syntactic features, which complicates the development of universal models. Unlike Modern Standard Arabic (MSA), dialects lack standardized orthography and often involve informal, colloquial expressions, making text normalization difficult. Additionally, frequent code-switching between dialects and MSA, as well as the use of borrowed words from other languages, further complicates sentiment detection. Limited annotated datasets and linguistic resources for many dialects restrict the training and evaluation of effective models. Moreover, sentiment analysis must consider cultural nuances, idiomatic expressions, and contextual meanings unique to each dialect to accurately capture the emotional tone of the text. These challenges necessitate specialized approaches and resources to improve sentiment analysis performance in dialectal Arabic.

In this research, we introduce a new dataset consisting of Saudi Arabian proverbs annotated with sentiment classifications. Additionally, we fine-tuned the CAMeLBERT-DA-SA model for sentiment analysis of Dialectal Arabic texts, utilizing both customer reviews and the newly created proverbs dataset. Our focus is on Saudi and Darija (Moroccan) dialects within the hospitality domain. Furthermore, we evaluated several open Arabic-centric large language models (LLMs) on the same domain, using the test set provided by the Ahasis

Proceedings of the Shared Task on Sentiment Analysis for Arabic Dialects (AHaSIS) associated with RANLP 2025, pages 69–75, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-109-7-011

shared task organisers.

2 Related Work

Recent advancements in Arabic sentiment and sarcasm analysis have increasingly adopted deep learning and ensemble-based methods to improve performance across various language tasks. (Gaanoun and Benelallam, 2021) introduced an ensemble framework that integrated Gaussian Naive Bayes, MARBERT, and a BiLSTM model enhanced with Mazajak embeddings. Their approach, fine-tuned using thecArSarcasm-v2 (Abu Farha et al., 2021) dataset and utilizing a weighted ensemble strategy, achieved improved accuracy in sarcasm detection. Similarly,(Wadhawan, 2021) explored two transformer-based models; AraBERT and AraELECTRA for both sentiment and sarcasm analysis. Evaluated on the ArSarcasm-v2 (Abu Farha et al., 2021) dataset with comprehensive pre-processing, the study found that AraBERT consistently outperformed AraELECTRA. Ensemble strategies continued to prove effective in the work of (Karfi and Fkihi, 2022), who employed CAMeLBERT, AraBERTv0.2, and a majority voting mechanism across multiple datasets, including MSA and dialectal Arabic sources. Their method, supported by thorough pre-processing, showed strong sentiment classification performance and demonstrated the benefits of combining model outputs. (Mohamed et al., 2022) further extended ensemble modeling by integrating the multilingual XLM-T model with the Arabic-centric MARBERT. To handle data imbalance issues, they applied focal loss and label smoothing techniques. Fine-tuning on datasets such as ASAD (Alharbi et al., 2021), SemEval-2017 (Rosenthal et al., 2017), and ArSarcasm-v2 (Abu Farha et al., 2021), their ensemble model outperformed all individual components, underscoring the strength of using hybrid multilingual-monolingual transformer-based models. In a broader evaluation, (Khondaker et al., 2023) assessed the capabilities of ChatGPT and GPT-4 in processing Arabic, comparing their performance in both Modern Standard Arabic (MSA) and Dialectal Arabic (DA). Their study spanned 44 natural language understanding and generation tasks across roughly 70 datasets, incorporating both automatic and human evaluations. The results highlighted that although ChatGPT exhibited strong performance in English, it was surpassed by smaller models fine-tuned for Arabic, especially in handling dialectal variation—an area where ChatGPT and GPT-4 showed notable limitations. Most recently, (Alosaimi et al., 2024) introduced a hybrid AraBERT-LSTM architecture that combines the contextual embedding capabilities of AraBERT with the sequential modeling strength of LSTM networks. Their work explored various embedding strategies, including CBOW, Skip-Gram, and AraBERT embeddings, and benchmarked the model against a range of traditional and deep learning algorithms. The hybrid model demonstrated exceptional results, achieving over 97 % overall accuracy and 90.40 % on the SS2030 dataset (Alyami and Olatunji, 2020), reinforcing the effectiveness of combining transformer-based embeddings with recurrent architectures for Arabic sentiment analysis.

3 Datasets

In the Ahasis Shared Task (Alharbi et al., 2025a,b), the organizers released both the training and test sets via Codabench[1] for use in model development and evaluation. Participants were also permitted to utilize any additional publicly available linguistic resources or corpora to enhance their model training. Table 1 presents the total number of Dialectal Arabic tweets and proverbs used in this study for training and testing purposes.

Data set	#Sentences
Train	1558
Test	216

Table 1: Number of Sentences in Train and Test sets

3.1 Ahasis Train Set

The training dataset (ATS)[2] is available in CSV format and contains a total of 860 hotel reviews, equally divided between 430 reviews in the Saudi dialect and 430 in Darija (Alharbi et al., 2025a,b). This dataset was assembled to assess the performance of Dialectal Arabic sentiment analysis tasks in the hospitality domain. Each record includes a unique identifier ("Original_ID"), a dialect label ("Dialect") specifying either Saudi or Darija, a sentiment classification ("Sentiment") with one of three values—Positive, Negative, or Neutral—and the review text ("Text").

[1] https://www.codabench.org/competitions/5871/

[2] https://drive.google.com/file/d/12PebNlUTrkpUb4B3s8GM6joiWp-bcgw5/view

3.2 Saudi Proverbs Dataset (SPD)

Participants in the Ahasis Shared Task are allowed to use any external resources and tools for training and fine-tuning. In addition to the official training set, we contribute a new dataset of 698 Saudi proverbs annotated with sentiment labels. The Saudi Proverbs Dataset (SPD) was developed to support research in Dialectal Arabic sentiment analysis with a focus on Saudi dialect proverbs. The methodology includes structured data collection, pre-processing, and sentiment annotation with a focus on reliability through inter-annotator agreement. SPD contains proverbs from the main sub-dialects of Saudi Arabic. Saudi Arabian sub-dialects include: Najdi, Hijazi, Shamali, Janoub and Sharqawi (Alahmari et al., 2024; Alahmari, 2025) These proverbs were sourced from various Saudi regions.

3.2.1 Data Collection

The dataset comprises Saudi proverbs collected from three primary sources to ensure diversity and authenticity:

- **Printed Books**: The primary reference, was the book by Abdulkareem Aljuhaiman [3] titled "Popular proverbs in the heart of the Arabian Peninsula",(1983). الأمثال الشعبية في قلب الجزيرة العربية.
- **Elderly Speakers**: Proverbs were collected through interviews with older Saudi individuals across different regions. This oral component helped capture traditional, under-documented proverbs with regional and dialectal characteristics.
- **Online Forums and Social Media**: These sources offered contemporary and colloquial proverbs actively used in informal settings, reflecting modern usage and regional variety.

3.2.2 Pre-processing

The collected proverbs were preprocessed to ensure consistency and usability by applying the following cleaning methods:
Deduplication: Redundant entries and slight variants were identified and consolidated.
Cleaning: Non-proverbial expressions and irrelevant content were manually filtered out.
Normalization: Light text normalization was applied, including the removal of diacritics, while preserving dialectal features and original spelling conventions.
Sentiment Annotation: Each proverb was manually annotated for sentiment polarity, using one of the following three labels: Positive: Expresses encouragement, praise, wisdom, or optimism. Negative: Expresses criticism, warning, disapproval, or pessimism. Neutral: Emotionally neutral or descriptive, without strong positive or negative sentiment. Three native Arabic-speaking annotators independently assigned sentiment labels to each proverb. All annotators were familiar with Saudi dialects and cultural nuances.

3.2.3 Inter-Annotator Agreement

Inter-annotator reliability was measured using Fleiss' Kappa (Fleiss, 1971), a statistical measure suitable for evaluating agreement among more than two annotators on categorical data. A total of N proverbs were independently annotated for sentiment polarity (Positive, Negative, Neutral) by three annotators. The Fleiss' Kappa score was calculated as **K = 0.85**, indicating **Almost Perfect** agreement according to the commonly used interpretation scale by (Fleiss, 1971). These results suggest that the annotation guidelines were clear and consistently applied across annotators.

SPD can be acceseeable from Github [4]. The Figures 1, 2 and 3 show examples of Saudi proverbs in SPD with their English translation and sentiment labels. Table 2 presents the sentiment label distribution across both the SPD and TS.

Proverb: الي ما يعرف الصقر يشويه

English: He who doesn't know the falcon cooks it.

Sentiment: Negative

Figure 1: Example (1) of Saudi Proverbs from SPD

[3] https://archive.org/details/0_20240129_20240129_1437

[4] https://github.com/SalwaAlahmari/Saudi-Proverbs-Dataset

Proverb: الزين زين لو قعد من النوم

English: The beautiful remains beautiful even when just out of bed.

Sentiment: Positive.

Figure 2: Example (2) of Saudi Proverbs from SPD

Proverb: وجع ساعة ولا وجع كل ساعة

English: Pain for an hour is better than pain every hour.

Sentiment: Neutral

Figure 3: Example (3) oof Saudi Proverbs from SPD

Dataset	#Positive	#Negative	#Neutral
SPD	99	511	88
TS	308	336	216

Table 2: Sentiment Labels in Train set and Saudi Proverbs Dataset

The test set[5] is provided in CSV format and comprises 216 hotel reviews, evenly divided between the Saudi and Darija dialects, with 108 reviews from each. This set was specifically curated to evaluate model performance on Dialectal Arabic sentiment analysis within the hospitality domain. Each entry contains a unique identifier ("ID"), a reference to the original review ("Original_ID"), a dialect label ("Dialect") indicating whether the text is in Saudi or Darija, and the full review text ("Text").

4 Methodology

In this section, we present our baseline Camelbert-da-sentiment model, describe the fine-tuning procedure, and discuss the optimization of hyper-parameters. In addition, we evaluate the performance of Arabic-centric LLMs with zero-shot learning mood and compare the results with our baseline. In this study will select three Arabic-centric Large Lnaguge Models : Allam, ACeGPT and Jais Models.

4.1 Models Selection

- **CAMeLBERT-DA-SA Model** (Inoue et al., 2021)model is a specialized model developed by fine-tuning the CAMeLBERT Dialectal Arabic (DA) base model. The fine-tuning process utilized several benchmark datasets, including ASTD, ArSAS, and SemEval, to optimize its performance on sentiment analysis tasks.

- **Allam** (Bari et al., 2025)is an autoregressive transformer model developed from scratch by the National Center for Artificial Intelligence at SDAIA. Its training involves two stages: first on a large English corpus, followed by a mixed Arabic-English dataset. We used ALLaM-7B-Instruct-preview variant.

- **ACeGPT** (Liang et al., 2024; Zhu et al., 2024) is a collection of fully fine-tuned generative text models specifically designed for the Arabic language. The version 2 of the 8-billion parameter model is based on Meta-LLaMA-3-8B. This model was developed collaboratively by researchers from King Abdullah University of Science and Technology (KAUST), the Chinese University of Hong Kong, Shenzhen (CUHKSZ), and the Shenzhen Research Institute of Big Data (SRIBD).We used AceGPT-v2-8B variant.

- **Jais** family includes bilingual English-Arabic LLMs optimized for Arabic, available in two types: trained from scratch and adaptively pre-trained from Llama-2. The collection features 20 models ranging from 590M to 70B parameters, trained on Arabic, English, and code data. All are instruction fine-tuned as text-to-text generative models and developed by Inception and Cerebras Systems (Inception, 2024). We used jais-family-13b variant.

All models used in our experiments were sourced from the Hugging Face repository[6]. The implementation and execution of our code were carried out using the PyTorch Transformers library[7]. Hyper-parameters were carefully selected to achieve optimal performance while minimizing training time.

[5] https://drive.google.com/file/d/1iRwoEIJ8dE2dYpml5v-OgQC6k3xRg2hV/view

[6] https://huggingface.co

[7] https://pytorch.org/hub/huggingface_pytorch-transformers/

Parameters	Values
learning_rate	5e-5
max_length	512
per_device_train_batch_size	8
per_device_eval_batch_size	8
num_train_epochs	2

Table 3: Hyper-parameters for fin-tuning CAMeLBERT-DA SA Model model

5 Results and Discussion

As the Table 4 shows, the fine-tuned CAMeLBERT-DA model achieved the highest F1 score of 75%, clearly outperforming the other evaluated models. This result demonstrates the advantage of task-specific fine-tuning, particularly when dealing with sentiment analysis in Dialectal Arabic within a focused domain.

In contrast, the remaining models—Allam, ACeGPT, and Jais—were used in a zero-shot setting without any fine-tuning on the task-specific data. Among them, Allam achieved the best performance with an F1 score of 70%, followed by ACeGPT at 68%, and Jais at 65%. Although these models showed reasonable performance, the results underline the limitations of applying even strong Arabic-centric LLMs directly to dialectal sentiment analysis without adaptation.

These findings underscore the importance of fine-tuning with domain-specific and dialect-relevant data—particularly in low-resource settings or linguistically diverse contexts such as Saudi Arabic and Darija. General-purpose models often struggle to capture the subtle linguistic and cultural nuances necessary for accurate sentiment classification in these dialects. The relatively low F1-scores observed in our experiments can be attributed to several factors. Chief among them is the limited availability of annotated corpora for Dialectal Arabic sentiment analysis, which typically involves three sentiment classes: Positive, Negative, and Neutral. Additionally, due to time and computational constraints, we were unable to perform a comprehensive evaluation of how different hyper-parameter settings might impact the performance of the fine-tuned CAMeLBERT-DA sentiment analysis model. These challenges, compounded by the intensive training requirements and high computational demands of large language models such as Allam, ACeGPT, and Jais, contribute to the difficulty of achieving higher performance in dialectal Arabic sentiment analysis tasks.

Model	F1-Score
Camelbert-da-sa	75%
Allam	70%
ACeGPT	68%
Jais	65%

Table 4: F1-Score of Selected Models on the test set

6 Conclusion and Future Work

In this study, we addressed the challenges of sentiment analysis in Dialectal Arabic by focusing on Saudi and Moroccan (Darija) dialects within the hospitality domain. We introduced a new dataset of Saudi Arabian proverbs annotated with sentiment labels and fine-tuned the CAMeLBERT-DA model using both customer reviews and proverbs. In addition, we evaluated three open-source Arabic-centric large language models—Allam, ACeGPT, and Jais—in a zero-shot setting using the Ahasis shared task test set. Our experimental results demonstrate that domain-specific fine-tuning significantly improves sentiment classification performance, as evidenced by CAMeLBERT-DA-SA achieving the highest F1 score. The results also highlight the limitations of zero-shot approaches for Dialectal Arabic sentiment analysis, even when using strong pre-trained LLMs.

For future work, we plan to explore few-shot and in-context learning methods to enhance zero-shot performance of large models on dialectal data. We also aim to expand our proverb dataset to include more dialects and develop more robust annotation guidelines. Finally, integrating multimodal sentiment cues (e.g., emojis, images) from social media could offer deeper insights into the sentiment expressed in dialectal Arabic contexts.

References

Ibrahim Abu Farha, Wajdi Zaghouani, and Walid Magdy. 2021. Overview of the wanlp 2021 shared task on sarcasm and sentiment detection in arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Latifa Al-Sulaiti and Eric Atwell. 2006. The design of a corpus of contemporary arabic. *International Journal of Corpus Linguistics*, 11(2):135–171.

Salwa Alahmari. 2025. SADSLyC: A corpus for saudi Arabian multi-dialect identification through song lyrics. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 38–43, Abu Dhabi, UAE. Association for Computational Linguistics.

Salwa Alahmari, Eric Atwell, and Mohammad r Alsalka. 2024. Saudi arabic multi-dialects identification in social media texts. In *Intelligent Computing*, pages 209–217, Cham. Springer Nature Switzerland.

Basma Alharbi, Hind Alamro, Manal Alshehri, Zuhair Khayyat, Manal Kalkatawi, Inji Ibrahim Jaber, and Xiangliang Zhang. 2021. Asad: A twitter-based benchmark arabic sentiment analysis dataset.

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)*.

Wael Alosaimi, Hager Saleh, Ali A. Hamzah, Nora El-Rashidy, Abdullah Alharb, Ahmed Elaraby, and Sherif Mostafa. 2024. Arabbert-lstm: improving arabic sentiment analysis based on transformer model and long short-term memory. *Frontiers in Artificial Intelligence*, Volume 7 - 2024.

Sarah N. Alyami and Sunday O. Olatunji. 2020. Application of support vector machine for arabic sentiment classification using twitter-based dataset. *Journal of Information & Knowledge Management*, 19(01):2040018.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham Abdullah Alyahya, Sultan AlRashed, Faisal Abdulrahman Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Saad Amin Hassan, Dr. Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, Ahmed Abdelali, Jeril Kuriakose, Abdalghani Abujabal, Nora Al-Twairesh, Areeb Alowisheq, and Haidar Khan. 2025. ALLam: Large language models for arabic and english. In *The Thirteenth International Conference on Learning Representations*.

Ali Farghaly and Khaled Shaalan. 2009. Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing*, 8(4).

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Kamel Gaanoun and Imade Benelallam. 2021. Sarcasm and sentiment detection in Arabic language a hybrid approach combining embeddings and rule-based features. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 351–356, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. 2021. Arabic natural language processing: An overview. *Journal of King Saud University - Computer and Information Sciences*, 33(5):497–507.

Inception. 2024. Jais family model card.

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*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Ikram El Karfi and Sanaa El Fkihi. 2022. An ensemble of arabic transformer-based models for arabic sentiment analysis. *International Journal of Advanced Computer Science and Applications*, 13(8).

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. GPTAraEval: A comprehensive evaluation of ChatGPT on Arabic NLP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 220–247, Singapore. Association for Computational Linguistics.

Juhao Liang, Zhenyang Cai, Jianqing Zhu, Huang Huang, Kewei Zong, Bang An, Mosen Alharthi, Juncai He, Lian Zhang, Haizhou Li, Benyou Wang, and Jinchao Xu. 2024. Alignment at pre-training! towards native alignment for arabic LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Omar Mohamed, Aly M Kassem, Ali Ashraf, Salma Jamal, and Ensaf Hussein Mohamed. 2022. An ensemble transformer-based model for arabic sentiment analysis. *Social Network Analysis and Mining*, 13(1):11.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Anshul Wadhawan. 2021. AraBERT and farasa segmentation based approach for sarcasm and sentiment detection in Arabic tweets. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 395–400, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Jianqing Zhu, Huang Huang, Zhihang Lin, Juhao Liang, Zhengyang Tang, Khalid Almubarak, Abdulmohsen Alharthik, Bang An, Juncai He, Xiangbo Wu, Fei Yu, Junying Chen, Zhuoheng Ma, Yuhao Du, He Zhang, Emad Alghamdi, Lian Zhang, Ruoyu Sun, Haizhou Li, and Jinchao Xu. 2024. Second language (arabic) acquisition of llms via progressive vocabulary expansion.

A Gemini-Based Model for Arabic Sentiment Analysis of Multi-Dialect Hotel Reviews: Ahasis Shared Task Submission

Mohammed A. H. Lubbad
Erciyes University
engmlubbad@gmail.com

Abstract

This paper presents a sentiment analysis model tailored for Arabic dialects in the hospitality domain, developed for the Ahasis Shared Task. Leveraging the Gemini Pro 1.5 language model, we address the challenges posed by the diversity of Arabic dialects—specifically Saudi and Moroccan Darija. Our method utilized the official Ahasis dataset comprising 3,000 hotel reviews. Through iterative benchmarking, dialect labeling, sarcasm detection, and prompt engineering, we adapted Gemini Pro 1.5 for the task. The final model achieved an F1-score of 0.7361 and ranked 10th on the competition leaderboard. This work shows that prompt engineering and domain adaptation of LLMs can mitigate challenges of dialectal variation, sarcasm, and resource scarcity in Arabic sentiment classification. Our contribution lies in the integration of dialect-specific prompt tuning with real-time batch inference, avoiding retraining. This approach, validated across 3,000 competition samples and 700 internal benchmarks, establishes a novel template for Arabic-domain sentiment pipelines.

1 Introduction

Arabic is a morphologically rich and sociolinguistically complex language, exhibiting strong diglossia between its formal variant (MSA) and a multitude of spoken dialects. These dialects can differ dramatically across regions in vocabulary, syntax, and even script usage. Consequently, building robust sentiment analysis models for Arabic is significantly more challenging than for languages with greater standardization (ElSayed et al., 2020; Zrigui et al., 2021).

With the tourism industry's digital transformation, understanding nuanced customer feedback in native dialects becomes crucial for service quality and competitive positioning. However, current sentiment models underperform on such real-world hospitality datasets, revealing an urgent gap.

While pre-trained models like AraBERT and CAMeL have advanced sentiment classification for MSA, their performance degrades when applied to dialect-rich, noisy, and context-sensitive content typical of social media or domain-specific reviews. Furthermore, most existing datasets lack sarcasm annotation or domain specificity, which impedes model accuracy on real-

Proceedings of the Shared Task on Sentiment Analysis for Arabic Dialects (AHaSIS) associated with RANLP 2025, pages 76–85, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-109-7-012

world texts.

The Ahasis Shared Task (Alharbi et al., 2025a) specifically targeted sentiment detection in Saudi and Moroccan (Darija) dialects within hotel reviews, a domain rich with nuanced emotional expressions and culturally embedded idioms. The broader context of evaluating Large Language Models on Arabic Dialect Sentiment Analysis has also been explored (Alharbi et al., 2025b). This paper documents our solution, which ranked among the top ten submissions, combining prompt engineering of Gemini Pro 1.5 with a domain-customized preprocessing and benchmarking strategy designed to overcome these real-world gaps.

2 Related Work

Arabic sentiment analysis has evolved from early lexicon-based systems (Abdul-Mageed and Diab, 2014) to modern deep learning and transformer-based approaches. Models like AraBERT (Antoun et al., 2020) have provided significant advancements by being pre-trained on large Arabic corpora. However, AraBERT and similar MSA-trained models often underperform on dialect-rich datasets. Hybrid systems such as AraBERT-LSTM and attention-integrated BiLSTM networks have shown state-of-the-art results in dialectal corpora, achieving over 97% accuracy on benchmark datasets (Serrano et al., 2024). Studies further emphasize the importance of not applying MSA-style stemming to dialectal text, particularly Moroccan Darija, where meaning is often embedded in surface forms (Matrane et al., 2024). Attention mechanisms and ensemble learning have emerged as potent tools for capturing context and sentiment nuances in Arabic dialects (Ombabi et al., 2024).

Notably, hospitality sentiment in Arabic dialects remains underexplored. While LLMs like GPT and Gemini are advancing multilingual NLP, few studies have benchmarked them in structured, low-resource domains, such as Arabic hotel reviews.

3 Data

3.1 Ahasis Dataset

The Ahasis Shared Task dataset (Alharbi et al., 2025a) provided annotated hotel reviews in Arabic, balanced across two dialects—Saudi and Darija (Moroccan). For the purpose of model training, we utilized the official Ahasis training set, which comprises **860 annotated reviews**. Each entry contains the review text, its dialect, and a sentiment label. The sentiment distribution of this training set is presented in Table 1. This distribution is notably imbalanced, with a significant proportion of negative samples and a smaller proportion of neutral samples compared to positive ones. The task demanded that participants train and test models capable of handling both dialect and sentiment classification under noisy, real-world conditions.

Sentiment	Count	% of Total
Negative	336	39.07%
Neutral	216	25.12%
Positive	308	35.81%

Table 1: Ahasis Training Set Sentiment Distribution (860 Samples)

3.2 Internal Benchmark

In addition to the official Ahasis dataset, we constructed a dedicated internal benchmark comprising **577 manually annotated YouTube comments** sourced from AJ360 shows, which focus on Arabic media content. This dataset was designed to simulate domain transfer challenges and evaluate model robustness for real-time sentiment detection in a less controlled, more colloquial environment.

The annotation of this internal dataset was performed in-house by a specialized data analytics team, requiring approximately **4-5 hours of fully focused and concentrated effort**. The original sentiment distribution of this benchmark (577 comments) was as follows:

- **Negative**: 55 samples ($\approx 9.53\%$)
- **Neutral**: 334 samples ($\approx 57.89\%$)
- **Positive**: 188 samples ($\approx 32.58\%$)

To mitigate the observed class imbalance and enhance model robustness, especially for minority classes (Negative and Neutral), simple data augmentation through **manual paraphrasing** was applied. This process expanded the dataset from its original 577 comments to a total of **700 comments**. The resulting sentiment distribution after augmentation, contributing to a slightly more balanced representation across sentiment categories for training purposes, is:

- **Negative**: 78 samples ($\approx 11.14\%$)
- **Neutral**: 274 samples ($\approx 39.14\%$)
- **Positive**: 348 samples ($\approx 49.71\%$)

This distribution, notably featuring a reduction in the majority of neutral comments and an increase in negative samples, reflects the nuanced and often ambiguous nature of sentiment in informal Arabic social media. We assessed candidate models using this benchmark before the final competition submission, providing crucial insights into their performance beyond the Ahasis-specific domain and aiding in early error analysis.

4 Methodology

4.1 Preprocessing

We designed a preprocessing pipeline to address the linguistic messiness inherent in social media and review texts, aiming to prepare the data for optimal large language model inference:

- **Cleaning**: Systematic removal of hyperlinks, user mentions (@mentions), emojis, and redundant whitespace.
- **Standardization**: Normalization of elongated words, e.g.(مراحب مراااااحب), and informal spellings in dialectal Arabic.
- **Dialect Tagging**: Automatic classification into Saudi vs. Moroccan Darija via dedicated language models; tags are injected into the prompt.
- **Sarcasm Flagging**: Combined ArSarcasm dataset (Alsarhan et al., 2021) with heuristic rules (e.g. contradiction patterns) to flag potential sarcasm.
- **Manual Verification**: Expert review of ambiguous/outlier cases to ensure data quality.

4.2 Prompt Engineering and Inference Setup

Our approach uses Google's hosted `gemini-1.5-pro` API, orchestrated via a spreadsheet Apps Script:

- Batch inference for large volumes of reviews.

- Few-shot JSON prompt with 20 dialect-balanced examples (see Appendix A).

- Output constrained to `positive`, `neutral`, or `negative`.

The API calls use:

- `temperature=0`

- `topP=0.95`

- `maxOutputTokens=8192`

Safety settings are all set to `BLOCK_NONE` to avoid filtering legitimate content.

The complete implementation of the batch inference script, including the detailed logic for API calls and result handling, is publicly available at: https://github.com/mlubbad/ahasis-sentiment-analysis.

4.3 Ahasis Training Set Sentiment Distribution

While not heavily skewed, this imbalance, particularly the smaller proportion of neutral samples, may have contributed to a tendency for the model to overpredict the majority classes, especially positive sentiment, as further discussed in the subsequent error analysis. Such distributional skew is critical to consider when evaluating model generalization, particularly in sentiment tasks where neutrality is often subtle and context-dependent.

4.4 Error Analysis and Prompt Refinement

Despite the strong performance on the Ahasis test set, a detailed analysis of misclassifications, particularly during the iterative prompt refinement process, provided crucial insights into the model's current limitations. We identified two primary categories of errors.

First, the model frequently overpredicted **positive** sentiment in **neutral** contexts, particularly when reviews contained polite or descriptive language that lacked explicit emotional cues. This suggests difficulty in distinguishing purely functional appreciation or factual statements from genuine positive sentiment. Examples illustrating this include:

- **True: Neutral | Predicted: Positive** الخدمة جيدة و موقع الفندق جيد قريب من المطار *("The service is good and the hotel location is good, close to the airport.")*
  → A purely factual statement about services and location was misinterpreted as expressing positive emotion.

- **True: Neutral | Predicted: Positive** رخيص وفي احسن موقع حصلت خصم... *("Cheap and in the best location, I got a discount...")*
  → The model incorrectly equated a statement of financial benefit with positive sentiment.

- **True: Neutral | Predicted: Positive** كننصح بالزيارة ديالو. الإنترنت مزيان... *("[I] recommend visiting it. The internet is good...") [Darija]*
  → A neutral recommendation in Moroccan Darija was over-interpreted as positive, highlighting challenges with dialect-specific expressions.

Second, the model struggled with reviews containing **implicit sentiment and sarcasm**. While our preprocessing pipeline included a sarcasm flagger, many instances rely heavily on cultural context and intricate linguistic nuances that are not easily captured by simple lexical cues or even explicit flagging. For example, a comment like الغرفة كانت رائعة لدرجة أنني لم أستطع النوم (*"The room was so wonderful I couldn't sleep"*) could be genuinely positive or highly sarcastic, a nuance the model often missed, typically defaulting to a literal (positive) interpretation. This pattern underscores a key challenge for LLMs in low-resource dialectal contexts where complex pragmatic understanding is required.

To mitigate these errors, we iteratively refined the prompt by adding curated examples of neutral, factual statements and ambiguous phrases to guide the model's understanding. While our chosen 'temperature' of 0 ensured deterministic outputs, which is beneficial for consistency, it limits the model's exploratory generation, potentially contributing to the observed bias. Future work will investigate strategies such as enhancing the prompt with more diverse and challenging neutral examples, exploring adaptive parameter tuning, and investigating post-processing techniques (e.g., calibrating output confidence thresholds for the 'neutral' class if API access allows) to rebalance predictions. This iterative prompt tuning process proved to be a practical method for targeted error correction in LLMs without requiring retraining.

5 Experiments & Results

5.1 Comparative Model Performance on Internal Benchmark

To rigorously assess Gemini Pro 1.5's capabilities and robustness prior to the Ahasis Shared Task submission, we conducted a comparative evaluation against a diverse set of ten transformer-based models on our **700-comment internal benchmark**. This benchmark, derived from manually annotated YouTube comments, was scaled proportionally to 700 to ensure consistent reporting and reflect the augmented dataset used for training. The models evaluated included prominent large language models such as GPT-4o, LLaMA-3, and Claude 3.5, as well as specialized fine-tuned regional models like CAMeL (Obeid et al., 2020) and AraBERT (Antoun et al., 2020). Evaluation metrics included macro-averaged F1-score and accuracy, complemented by confusion matrix analysis to assess class-wise behavior.

Table 2 presents the comparative results. On our internal benchmark, Gemini Pro 1.5 achieved the highest accuracy of 81.46% and a Macro-F1 score of 0.801, significantly outperforming all other tested models, including GPT-4o and LLaMA-3. The 95% confidence interval for Gemini Pro 1.5's Macro-F1 score on this dataset was determined to be $[0.6962, 0.7874]$.

This benchmark reinforced the selection of Gemini Pro 1.5 for the Ahasis submission, as it significantly outperformed other models, particularly in detecting the nuanced neutral sentiment, which is typically prone to misclassification in real-world social media data. The consistent superior performance on our internal benchmark, coupled with insights from confusion matrix analysis, provided crucial understanding of the model's strengths and areas for prompt refinement before the final competition submission.

To provide a more granular view of the model's performance and error patterns on the internal benchmark test set, Figure 1 presents the confusion matrix:

Model	Accuracy	F1-score
Gemini Pro 1.5	81.46%	0.801
GPT-4o	70.54%	0.692
LLaMA-3	70.36%	0.688
Claude 3.5	65.51%	0.641
GPT-4	64.47%	0.623
CAMeL	54.42%	0.498

Table 2: Comparative Model Performance Results on Internal Benchmark

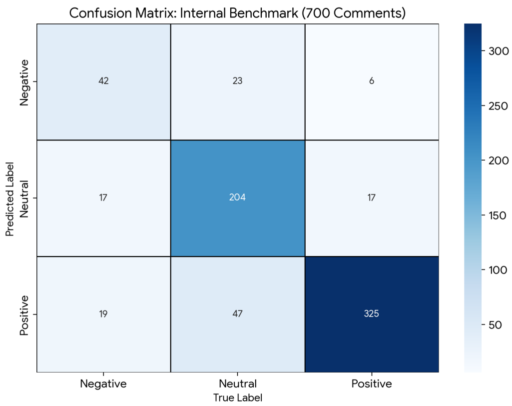


Figure 1: Confusion Matrix for Internal Benchmark (700 Samples)

5.2 Ahasis Submission Metrics and Confusion Analysis

The Ahasis Shared Task focused exclusively on sentiment analysis in Arabic hotel reviews from Saudi and Moroccan (Darija) dialects. Unlike media-based sentiment, which often skews toward polarized opinion, hospitality reviews frequently contain nuanced, mixed sentiments and indirect criticism. The Ahasis dataset posed a realistic challenge due to its domain specificity, balanced sentiment classes, and dialectal variance, making it a strong benchmark for testing robustness in real-world sentiment systems.

Metric	Value
F1-score	0.7361
Accuracy	0.7361
Precision	0.7361
Recall	0.7361
Balanced Accuracy	0.7229

Table 3: Leaderboard results on Ahasis test set

These results, **directly obtained from the official Ahasis leaderboard**, place our submission among the top-performing entries, affirming that prompt-engineered large language models like Gemini Pro 1.5 can effectively handle Arabic sentiment classification in niche domains. The identical values across F1-score, accuracy, precision, and recall, alongside a balanced accuracy of 0.7229, indicate a consistent and robust performance that effectively handles potential class imbalance and sentiment distribution skew, particularly in subtle neutral cases. This showcases the effectiveness of dialect-specific prompt tuning and heuristic preprocessing in addressing the challenges of domain-limited, dialect-rich data.

It is important to note that a direct statistical significance test, such as McNemar's test, comparing our model's performance on the Ahasis shared task against other baselines was not feasible, as the true labels for the Ahasis test set and the predicted labels from other participants were not made available to us.

While a detailed confusion matrix for the Ahasis test set is not publicly available for comparison, qualitative analysis of the model's performance, consistent with observations in Section 4.4, suggests ongoing challenges with the neutral class. The model tends to overpredict positive sentiment for subtle or ambiguous neutral texts, indicating a 'positive drift.' Similarly, some negative samples may also be incorrectly predicted as positive, and neutral samples as negative, reflecting the inherent complexities of informal Arabic sentiment. These patterns align with our error analysis findings and highlight areas for future prompt refinement.

6 Deployment

The selected model was integrated into a dashboard system within AJ360's media monitoring platform. Real-time analysis of social media comments (TikTok, YouTube, X, Facebook, Instagram) enabled the team to:

- Detect spikes in audience negativity during controversial broadcasts
- Compare sentiment shifts across platforms
- Generate weekly brand engagement summaries segmented by sentiment and dialect

The deployment used a REST API interface to connect the sentiment engine to AJ360's front-end interface, ensuring smooth scalability and operational use.

7 Discussion

Our results demonstrate that a large language model, guided by dialect-aware prompt engineering, can achieve competitive performance in a niche sentiment analysis task without task-specific fine-tuning. The model's 10th-place rank in the Ahasis shared task validates this prompt-centric approach as a viable strategy for low-resource dialectal domains.

The primary challenge remains the correct classification of the neutral class, a finding consistent with the broader sentiment analysis literature. Our error analysis (Section 4.4) revealed that this difficulty stems from two specific sources: the model's tendency to misinterpret factual descriptions of service quality as positive sentiment, and its failure to consistently detect culturally-nuanced sarcasm. This highlights that while LLMs possess vast world knowledge, their grasp of implicit, context-dependent sentiment in specific dialects is still limited.

To address these issues, future research should move beyond generic data augmentation. We propose exploring targeted strategies such as prompt-level augmentation, where the few-shot examples are dynamically weighted to include more challenging neutral and sarcastic cases, directly counteracting the positive skew noted in our dataset (Table 1). Furthermore, integrating semi-supervised techniques specifically for sarcasm labeling could prove more effective than relying on pre-existing, out-of-domain datasets.

It is important to acknowledge the limitations of this study. Our results are based on a single experimental run; therefore, future work should incorporate bootstrapping to establish confidence intervals, providing a more robust measure of performance variance. Additionally, while our findings validate that abstaining from MSA-style normalization (e.g., stemming) enhances performance on dialect-heavy texts, this conclusion should be further tested across a wider range of Arabic dialects. Visualizing attention weights, as suggested in prior work, could also offer greater interpretability into how the model processes dialectal versus MSA features.

8 Conclusion

This work demonstrates a high-performing sentiment analysis pipeline tailored to Arabic dialects. It achieved competitive performance in the Ahasis Shared Task and proved robust in real-world deployment. Our approach shows that dialect-informed preprocessing, benchmark-led model selection, and strategic fine-tuning of large models like Gemini Pro 1.5 yield impactful results. Future work will explore transfer learning across dialects, interpretability improvements, and integration of external knowledge sources (e.g., cultural ontologies).

References

Muhammad Abdul-Mageed and Mona Diab. 2014. SAMAR: Subjectivity and Sentiment Analysis for Arabic Social Media. *Computer Speech & Language*, 28(1):20–37.

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)*, Varna, Bulgaria. Association for Computational Linguistics (ACL).

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)*, Varna, Bulgaria. Association for Computational Linguistics (ACL).

Fawaz Alsarhan, Saad Albakr, and Abdulaziz Alodhayb. 2021. ArSarcasm: An Arabic Sarcasm Detection Dataset and Framework. *IEEE Access*, 9:158186–158197.

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 Resources Association (ELRA).

Alaa ElSayed, Ahmed Shaar, and Wajdi Zaghouani. 2020. Arabic Language Resources and Processing Tools for Sentiment Analysis: Current State and Future Directions. *Journal of Information Science*, 46(6):788–805.

Yassir Matrane, Faouzia Benabbou, and Zineb Ellaky. 2024. Enhancing Moroccan Dialect Sentiment Analysis through Optimized Preprocessing and Transfer Learning Techniques. *IEEE Access*, 12:16276–16298.

Osama Obeid, Salam Khalifa, Dana Abdulrahim, and Nizar Habash. 2020. CAMeL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Abubakr H. Ombabi, Wael Ouarda, and Adel M. Alimi. 2024. Improving Arabic Sentiment Analysis across Context-Aware Attention Deep Model. *Language*

Resources and Evaluation, 59(2):639–663.

Martín Serrano, Hager Saleh, Ali A. Hamzah, et al. 2024. ArabBERT-LSTM: Improving Arabic Sentiment Analysis based on Transformer Model and LSTM. *Frontiers in Artificial Intelligence*, 7:1408845.

Mohamed Zrigui, Haifa Ben Aicha, and Lamia Hadrich Belguith. 2021. Survey on Arabic Sentiment Analysis: Techniques, Resources and Challenges. *Artificial Intelligence Review*, 54(6):4271–4312.

Appendix A: Full Prompt Template

The following is the complete 20-shot prompt template used for guiding the Gemini Pro 1.5 model for sentiment analysis of Arabic hotel reviews. The prompt begins with a detailed persona and task definition, followed by specific guidelines and dialect-specific few-shot examples (represented here by the first example and its structure, with the understanding that 19 additional examples would follow the same pattern).

```
You are a professional data scientist and
    NLP specialist with extensive
    experience in sentiment analysis,
    particularly in Arabic dialects. Your
    primary task is to classify the **
    overall sentiment** of **Arabic hotel
    reviews** into one of three categories:
     **positive**, **neutral**, or **
    negative**.

            Arabic presents unique
                challenges due to its rich
                variety of dialects beyond
                Modern Standard Arabic (MSA
                ). Each -dialectsuch as
                Saudi Arabic and -Darijacan
                 significantly differ in
                vocabulary, syntax, and
                idiomatic expression,
                especially in informal
                reviews. Your analysis must
                 handle these linguistic
                variances accurately.

            ##   Task Definition

            **Classify the sentiment of
                Arabic hotel review texts**
                 into:
            - 'positive
            - 'neutral'
            - 'negative'

            ##   Dataset Structure

            Each review is labeled with:
            - **Text**: The Arabic review
                text.
            - **Sentiment**: The ground-
                truth sentiment label (
                positive, negative, or
                neutral).
            - **Dialect**: The regional
                variant of Arabic (e.g., '
                Saudi', 'Darija').

            ##   Guidelines

            - Strict to trained data first
                while classifying not to
                your knowledge.
            - Focus exclusively on the **
                overall sentiment**
                expressed by the reviewer,
                not isolated phrases.
            - Prioritize dialect-specific
                nuances and idiomatic
                expressions (e.g., sarcasm,
                 exaggeration).
            - **Do not** infer sentiment
                from commands or meta-
                commentary in the review (e
                .g., "please fix the air
                "conditioning  Negative
                unless frustration is
                clearly implied).
            - If an example is available
                and matches the pattern,
                use that **as a benchmark
                **.
            - Avoid literal translation or
                relying on formal Arabic
                sentiment if dialectal cues
                 suggest a different tone.
            - Output **only the sentiment
                label**: Positive, Neutral,
                 or Negative.
            - **Do not** explain your
                answer or add any
                commentary.
            Let us start
            Dialect: Saudi, Text:
```

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.

Author Index