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.

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

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

[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.

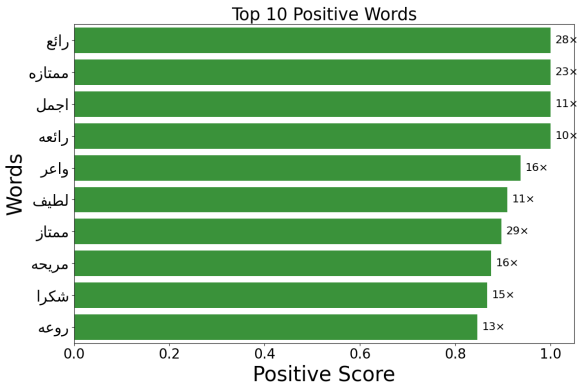

Figure 1: Top 10 Positive Words by Frequency in the Arabic Sentiment Lexicon

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

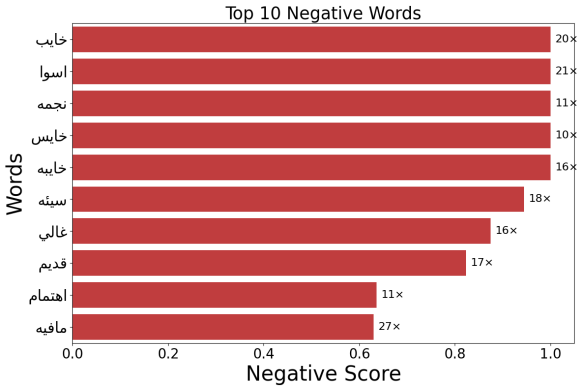

Figure 2: Top 10 Negative Words by Frequency in the Arabic Sentiment Lexicon

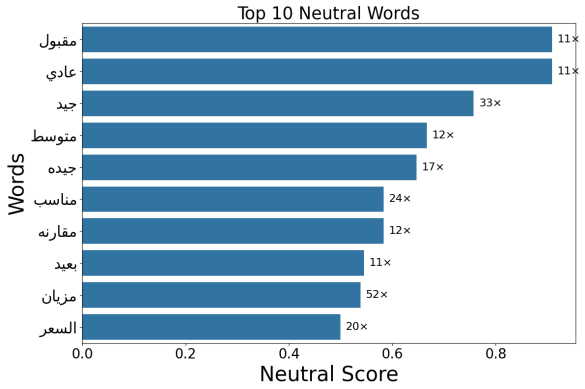

Figure 3: Top 10 Neutral Words by Frequency in the Arabic Sentiment Lexicon

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

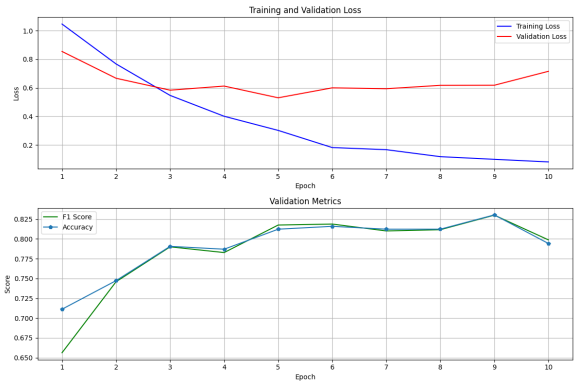

Figure 4: Training and validation performance over 10 epochs for our hybrid model

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.