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 hyperparameters. 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.

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

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 pretrained 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]. Hyperparameters were carefully selected to achieve optimal performance while minimizing training time.

[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.