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.

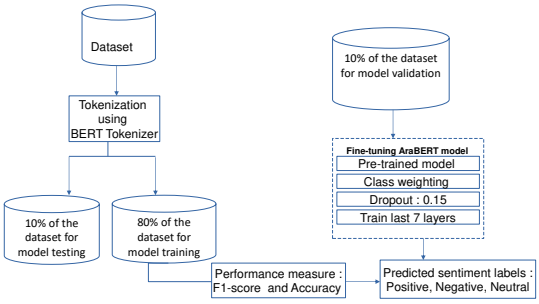


Figure 1: The system architecture.

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.	darija	Negative
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.)		
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.

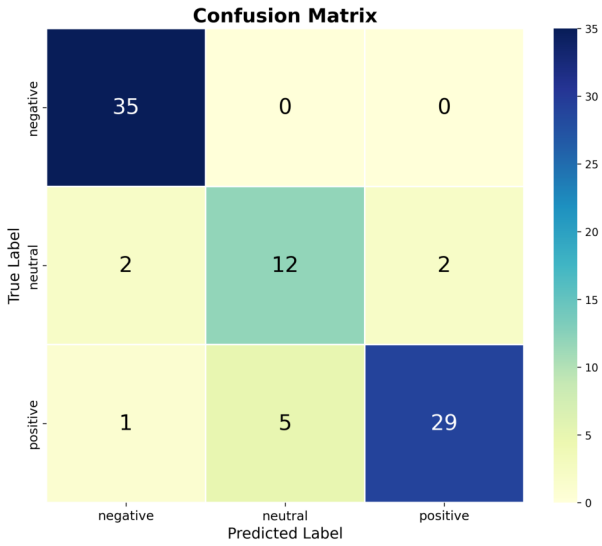


Figure 2: Confusion matrix

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 overreliance 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.

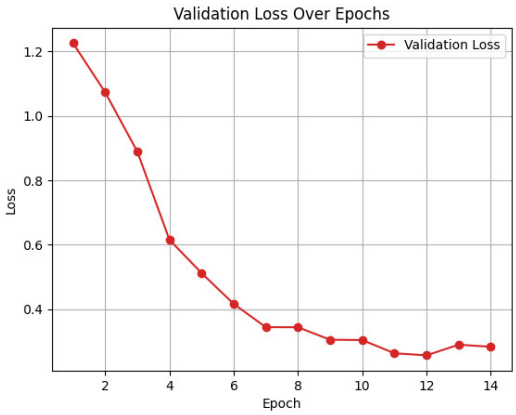
The model attained an F1-score of 0.88 on the standard test set, reflecting a modest 5% decline from the validation score (0.93). This marginal drop is consistent with typical generalization behavior, suggesting that the model performs robustly on data sampled from a similar distribution as the training 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:

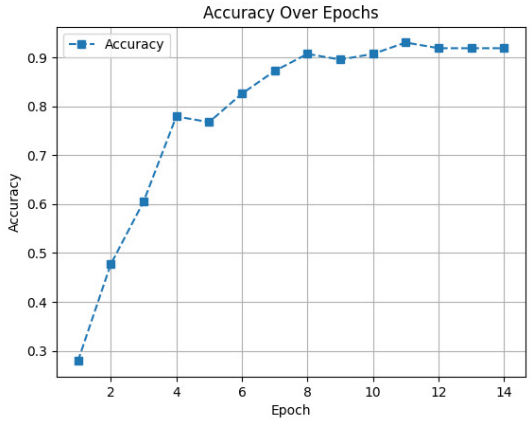
- Distributional differences between the blind test data and the training/validation sets, possibly due to unseen variations or domain shifts.
- Limited generalizability of some learned patterns, implying that the model may rely on features that do not transfer effectively to entirely new data.
- Potential 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:

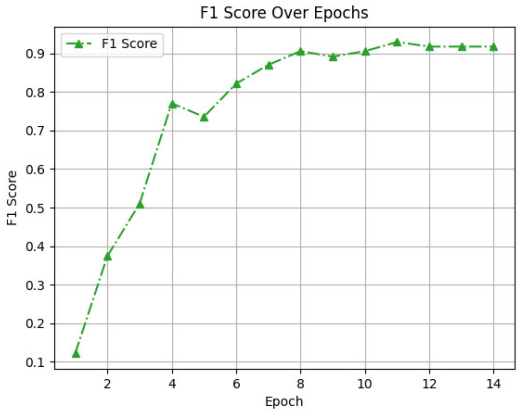
- Data augmentation and domain adaptation to improve generalization across diverse scenarios.
- Expanded dataset collection, particularly targeting underrepresented or edge cases to reduce distributional biases.
- Further analysis of feature representations to identify and mitigate non-transferable learned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