Evaluating Large Language Models on Sentiment Analysis in Arabic Dialects

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

Abstract

Despite recent progress in large language models (LLMs), their performance on Arabic dialects remains underexplored, particularly in the context of sentiment analysis. This study presents a comparative evaluation of three LLMs, DeepSeek-R1, Qwen2.5, and LLaMA-3, on sentiment classification across Modern Standard Arabic (MSA), Saudi dialect and Darija. We construct a balanced sentiment dataset by translating and validating MSA hotel reviews into Saudi dialect and Darija. Using parameter-efficient fine-tuning (LoRA) and dialect-specific prompts, we assess each model under matched and mismatched prompting conditions. Experimental results show that Qwen2.5 achieves the highest macro F1 score of 79% on Darija input using MSA prompts, while DeepSeek performs best when prompted in the input dialect, reaching 71% on Saudi dialect. LLaMA-3 exhibits stable performance across prompt variations, with 75% macro F1 on Darija input under MSA prompting. Dialect-aware prompting consistently improves classification accuracy, particularly for neutral and negative sentiment classes.

1 Introduction

Sentiment Analysis (SA) is a fundamental task in Natural Language Processing (NLP) that involves the computational identification and categorisation of emotions, opinions, and attitudes expressed in text (Wankhade et al., 2022; Birjali et al., 2021). Recent advancements in transformer-based architectures and Large Language Models (LLMs) have significantly enhanced SA performance by enabling deeper contextual understanding and improved generalisation across varied text inputs (Zhang et al., 2024; Krugmann and Hartmann, 2024).

Among the various application areas of SA, the hospitality domain has proven particularly valuable due to the abundance of user-generated reviews, which offer rich insights into customer experiences. Hotel reviews encapsulate sentiment-rich narratives that directly influence consumer behaviour and business strategies. This makes hospitality a high-impact domain for testing and refining SA techniques (Ameur et al., 2023). Similarly, SA in the Arabic language is gaining prominence due to the expanding digital presence of Arabic-speaking populations. However, this progress is uneven as Modern Standard Arabic (MSA) has been the primary focus, while dialectal varieties, which dominate informal communication, remain under-represented in both research and resources (Khaled et al., 2024; Sherif et al., 2023).

Despite the growing interest in Arabic NLP, dialectal SA remains under-explored. The lack of annotated datasets, along with the linguistic diversity and informal nature of dialects, complicates model development and evaluation (Alotaibi and Nadeem, 2024). Furthermore, phenomena such as orthographic ambiguity and regional lexical variation reduce the effectiveness of models trained primarily on MSA. These limitations underscore the need for adaptive approaches that can generalise across dialects without requiring extensive labelled corpora (Balakrishnan et al., 2025).

To address these challenges, this study investigates the use of LLMs for sentiment classification across Arabic dialects, specifically focusing on Saudi dialect, Darija and MSA. By leveraging prompt-based learning, the proposed framework enables flexible and domain-adaptive sentiment classification. Inspired by recent findings that highlight the role of prompt engineering in improving LLM performance (Rossyaykin, 2024), our approach incorporates iterative prompt refinement and example selection techniques to better handle dialectal variance.

The rest of this paper is structured as follows:

Proceedings of Recent Advances in Natural Language Processing,pages 67–74
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-008

Section 2 reviews prior research on Arabic sentiment analysis. Section 3 details the proposed methodology. Section 4 presents the experimental results and evaluation. Section 5 discusses our findings. Finally, Section 6 concludes the paper and outlines future research directions.

2 Related Work

Arabic SA presents unique challenges due to the linguistic complexity, rich morphology, and wide dialectal variation across regions. These difficulties are further compounded by the lack of large, high-quality annotated resources. Studies such as (Khaled et al., 2024) highlight the need for more datasets and fine-tuned models to bridge this gap.

Miah et al. (2024) proposed a resource-agnostic approaches that tackled the challenge of performing SA on non-English texts, including Arabic, by employing a cross-lingual framework. Their methodology involved translating Arabic texts into English using Google Translate and LibreTranslate, followed by sentiment classification using an ensemble of pre-trained English-language models,Twitter-RoBERTa, multilingual BERT and GPT-3. The final sentiment label was derived via majority voting. This setup yielded up to 86.71% accuracy for Arabic inputs.

Abuhammad and Ahmed (2024) developed a more linguistically grounded approach by focusing on negation detection, a key challenge in Arabic sentiment classification due to its ability to reverse sentiment polarity. They compiled a dataset of 84,000 Arabic hotel reviews, evenly split between positive and negated-positive examples, and engineered hybrid feature sets combining lexical elements and structural cues. The Deep Learning classifier achieved a 99.24% accuracy, outperforming all traditional models.

Expanding the scope to dialect-specific evaluations, Qarah and Alsanoosy (2025) conducted a large-scale benchmarking of BERT-based Arabic models, focusing particularly on Moroccan Arabic (Darija). They evaluated 14 pre-trained transformer models, including multi-dialect (e.g., MARBERTv2, QARiB), non-Moroccan mono-dialect (e.g., SaudiBERT, EgyBERT), and Darija-specific models (e.g., DarijaBERT variants, MorRoBERTa). Each model was fine-tuned and tested under a uniform pipeline across 13 NLP tasks and 11 datasets, ensuring methodological rigor and reproducibility. Evaluation metrics included F1-score and accuracy, with the best result per task reported.

Shifting to generative models,Al-Thubaity et al. (2023) evaluated GPT-3.5, GPT-4, and Bard AI on Saudi Dialect Arabic sentiment tasks using the Saudi Dialect Twitter Corpus (SDTC). The LLMs were benchmarked against fine-tuned BERT models, with GPT-4 achieving 77% F1-score, close to the top performing BERT model 79%. While LLMs excelled in negative sentiment detection, neutral sentiment remained a challenge. They also assessed LLM-generated tweets for data augmentation, but found that these did not enhance BERT performance, suggesting limitations in synthetic data realism. Finally, the most recent approach of Zouidine and Khalil (2025) investigates the effectiveness of general-purpose open-source LLMs (LLaMA, Mixtral, and Gemma) for sentiment classification in MSA. The models were instructed in English to classify the Arabic reviews as positive or negative. With a small set of labeled examples, LLaMA 3 achieved 84.8% accuracy in the 3-shot setting, nearly matching the AraBERTv2 baseline 87%. This shows that LLMs can approximate task-specific models when guided appropriately, though they still fall slightly short of dedicated, fine-tuned architectures. The findings affirm the potential of prompt-based learning in Arabic sentiment tasks, especially in settings with limited labeled data.

While prior research has demonstrated the potential of transformer-based models, cross-lingual setups, and prompt-based LLMs for Arabic sentiment analysis, key gaps remain unaddressed. Dialect-specific evaluations are often constrained by the lack of balanced, parallel datasets, making fair comparisons across varieties difficult. In addition, most existing datasets are heavily skewed toward MSA, limiting their representativeness of real-world dialectal usage. Furthermore, there has been limited systematic investigation into how prompt–input alignment influences model behaviour across Arabic dialects. These limitations underscore the need for controlled, comparative studies that account for dialectal variation and prompt design.

3 Methodology

We conducted experiments using three large language models: Qwen2.5-7B-Instruct, LLaMA3 and DeepSeek-R1. All models were evaluated under the same experimental setup. Sentiment labels were standardised and used to create an 80/20 stratified train/test split for each dialect to maintain

balanced class distribution.

3.1 Data

For this study, we used the ABSA-Hotels dataset, released as part of the Arabic track of SemEval-2016. The dataset consists of Arabic hotel reviews collected from popular platforms such as Booking.com and TripAdvisor. (Pontiki et al., 2016; Mohammad et al., 2016; Al-Smadi et al., 2019).

We carried out a thorough cleaning and restructuring process. Sentences with conflicting or mixed polarity labels were manually reviewed and reassigned using consistent heuristics. Text normalisation was applied using the ruqiya library, removing punctuation, diacritics and elongated characters. Duplicates and very short entries were discarded. From this process, we constructed a balanced subset of 538 sentences. To extend the dataset's across dialectal Arabic, we translated MSA sentences into Saudi dialect and Darija using Meta's NLLB-200 translation model. Each translation was manually evaluated and corrected where necessary, with validation performed by native speakers to ensure dialect accuracy, translation correctness, and sentiment preservation.

3.2 Experimental Setup

3.2.1 Prompt Design

The prompt design followed a structured, instruction-based format using explicit system, user and assistant role tags, consistent with chat-style LLM interfaces.

We used dialect-specific prompts tailored to each Arabic variety, ensuring that the instructions and examples reflected the linguistic features of the target dialect. Additionally, we tested both MSA and English prompts across all input dialects for comparative analysis.

Each prompt consisted of System instruction that defined the assistant's role as a sentiment classification expert for a specific dialect and emphasised that the result must be a direct label. Then, user query presented the input sentence for classification. Finally, assistant response for the expected model output was the correct sentiment label only, without any explanation or justification.

A separate template was constructed for each dialect, adapting instruction tone and vocabulary to the dialect's characteristics. Table 1 presents sample prompts across dialects.

Additionally, three short reviews were randomly sampled from each dialect-specific dataset, one per sentiment class, and used as few-shot exemplars in the prompt construction.

3.2.2 Model Fine-Tuning

Fine-tuning was performed using parameter-efficient fine-tuning (PEFT) via Low-Rank Adaptation (LoRA), which updates a limited set of trainable parameters while keeping the base model weights frozen (Hu et al., 2021).

LoRA adapters were integrated into the attention projection layers, specifically those responsible for generating the query, key, and value vectors. The output projection layer, which maps the attended values back to the original embedding space, was also involved.

To further improve efficiency and performance, we applied 4-bit quantization during fine-tuning. This reduced memory usage, also showed improved performance across the evaluated tasks.

Training was conducted using the Hugging Face Trainer API. Model evaluation was performed at the end of each epoch using a held-out test set, and the best-performing checkpoint was selected based on evaluation loss. All model weights, tokenizer configurations, and training logs were saved to ensure full reproducibility. The final fine-tuned model was subsequently used for downstream inference and analysis.

4 Evaluation Results

This section presents the evaluation results of three LLMs (DeepSeek, Qwen2.5, and LLaMA-3) on sentiment classification tasks across different Arabic Varieties. We analyse their performance under various prompting strategies, including dialect-matched, MSA, and English prompts, to assess how prompt–input alignment impacts classification accuracy and macro F1 scores.

4.1 Effect of Dialectal Prompting on Model Performance

Table 2 highlights DeepSeek's performance across different prompt–dialect configurations. The model shows a clear advantage when the prompt language is aligned with the dialectal variety of the input. Its best results were obtained on Saudi data using a Saudi dialect prompt and Darija data using a Darija prompt, with accuracies of 72% and 66%, and macro F1-scores of 71% and 64%, respectively. On MSA data, DeepSeek performed comparably with both English and MSA prompts, achieving

Prompt Language/Dialect	System Instruction	Few-Shot Examples
English	You are an intelligent assistant skilled in understanding Arabic and analyzing sentiment with high accuracy. Your task is to read the sentence and classify the sentiment it expresses into only one of the following categories: "positive", "negative", or "neutral". The answer must be precise, clear, and without explanation. What is the type of sentiment?	Sentence: "أحب هذا المنتج كثيرًا" Sentiment: positive Sentence: "الخدمة كانت سيئة جدًا" Sentiment: negative Sentence: "المنتج عادي، لا بأس به" Sentiment: neutral
MSA	أنت مساعد ذكي متمرس في فهم اللغة العربية وتحليل المشاعر بدقة عالية. مهمتك هي قراءة الجملة وتصنيف المشاعر التي تعبّر عنها إلى واحدة فقط من الفئات التالية: "إيجابي"، "سلبي"، أو "محايد". يجب أن تكون الإجابة دقيقة وواضحة و بدون شرح الرجاء تحليل المشاعر في الجملة التالية: ما هو نوع المشاعر؟.	العبارة: 'أحب هذا المنتج كثيرًا'صنف المشاعر إلى: إيجابي العبارة: 'الخدمة كانت سيئة جدًا'صنف المشاعر إلى: سلبي العبارة: 'المنتج عادي، لا بأس به' صنف المشاعر إلى: محايد
Saudi	أنت مساعد ذكي و تفهم في اللهجة العربية السعودية وتحليل المشاعر بدقة عالية. مهمتك هي انك تقرأ الجملة وتصنف المشاعر اللي تعبّر عنها إلى وحدة بس من هذي الفئات: "إيجابي"، "سلبي"، أو "محايد". لازم تكون الإجابة دقيقة وواضحة وبدون شرح حلل المشاعر في الجملة التالية: وش هو نوع المشاعر؟.	العبارة: 'أحب هالفندق مره' صنف المشاعر إلى: إيجابي العبارة: 'الخدمة كانت مره سيئة 'صنف المشاعر إلى: سلبي العبارة: 'الخدمة عادية، ما عليها' صنف المشاعر إلى: محايد
Darija	اعتبر راسك مساعد دكي، كتفهم العربية مزيان و كتعرف تدير تحليل المشاعر بدقة كبيرة. المهمة ديالك هي تقرا الجملة و تعطي الشعور و اللي كتعبر عليه، اختار شعور واحد من هادو "محايد" او "سلبي"، او "ايجابي"، و تكون الاجابة دقيقة و دون شرح عفاك حاول تعرف الاحساس اللي كتعبر عليه هاد الجمل: شنو هو نوع الشعور؟.	العبارة : 'كنبغي هاد المنتج بزاف' صنف المشاعر إلى : إيجابي العبارة : 'السيرفيس كان خايب بزاف' صنف المشاعر إلى : سلبي العبارة : 'المنتج عادي، مابيهش' صنف المشاعر إلى : محايد

Table 1: Prompt templates per language/dialect

Data	Prompt Language	Accuracy	Macro F1
Darija	English	48%	40%
	MSA	64%	60%
	Darija	**66%**	64%
Saudi	English	67%	51%
	MSA	65%	54%
	Saudi	**72%**	**71%**
MSA	English	**70%**	67%
	MSA	**70%**	**69%**

Table 2: DeepSeek model performance across dialects and prompt types

Data	Prompt Language	Accuracy	Macro F1
Darija	English	31%	24%
	MSA	**81%**	**79%**
	Darija	78%	75%
Saudi	English	32%	27%
	MSA	**76%**	**74%**
	Saudi	66%	64%
MSA	English	33%	28%
	MSA	**74%**	**73%**

Table 3: Qwen2.5 performance across dialects and prompt types

70% accuracy in both cases. The macro F1-score was slightly higher with the MSA prompt at 69%, compared to 67% with the English prompt. In contrast, performance declined with MSA and English prompts, especially on Darija input, where the F1 dropped to 40% under English prompting.

A similar pattern, though more robust overall, is observed with Qwen2.5, as shown in Table 3. Among all models evaluated, Qwen2.5 achieved the strongest performance. On Darija data, it reached a macro F1-score of 79% with MSA prompts and remained strong under dialectal prompting, achieving 75%. On Saudi input, the model again benefited from MSA prompts, obtaining an F1-score of 74%, outperforming both LLaMA-3 and DeepSeek in that configuration. However, Qwen2.5 struggled with English prompts, where performance dropped significantly to 27% on Saudi data and 24% on Darija data.

In comparison, LLaMA-3 demonstrates a more balanced, though slightly less competitive, performance across configurations, as shown in Table 4. Unlike Qwen2.5, LLaMA-3 showed greater resilience to English prompts, maintaining F1-scores between 66% and 71% on both Darija and Saudi data. Its strongest results were observed with MSA prompts on Darija and MSA input, reaching macro F1-scores of 75% and 73%, respectively. On Saudi data, LLaMA achieved a solid 72% F1 with MSA prompts and 69% with English prompts, further

Data	Prompt Language	Accuracy	Macro F1
Darija	English	65%	66%
	MSA	**75%**	**75%**
	Darija	68%	58%
Saudi	English	69%	69%
	MSA	**75%**	**72%**
	Saudi	73%	67%
MSA	English	70%	71%
	MSA	**74%**	**73%**

Table 4: LLaMA-3 performance across dialects and prompt types

underscoring its cross-lingual flexibility. While dialectal prompting did improve performance in some cases, particularly on Saudi input, the results were mixed on Darija, suggesting that LLaMA-3 maintains stable yet prompt-sensitive behaviour and benefits most from formal or multilingual-aware prompting across Arabic varieties.

4.2 Impact of Dialectal Prompts on Neutral and Negative Sentiment

Dialect-specific prompts significantly improved performance on neutral and negative sentiment classes, which are especially difficult to classify in dialectal text (Al-Thubaity et al., 2023).

While macro F1 scores for the positive class remained relatively stable across prompts, neutral and negative predictions improved in several configurations when dialectal prompting was applied. However, the extent of improvement varied across models and dialects. In some cases, performance even declined slightly, particularly for Qwen2.5 on Darija input and LLaMA-3 on Saudi data, indicating that dialectal prompting is not consistently beneficial, but context-dependent.

Table 5 presents the neutral and negative F1 scores before and after applying dialect-specific prompts for each model and dialect. Notably, DeepSeek showed consistent improvement across nearly all settings, especially for the neutral class, rising from 13% to 55% on Saudi data. LLaMA-3 achieved its best negative F1 on Darija data when switching from MSA to Darija (77% to 88%), while Qwen2.5 remained strong across the board but saw minimal gain from dialect adaptation in some settings.

Overall, these findings suggest that dialectal prompts can help models better interpret challenging sentiment classes, particularly when dealing with informal input. However, they also highlight that not all models respond equally to prompt shifts, and its impact varies depending on the model's handling of linguistic variation and the specific features of the input.

5 Discussion

Building on the experimental results, this section discusses the implications of prompt language alignment, model-specific behaviour, and class-level performance patterns. We highlight differences in model sensitivity to dialectal cues, challenges in neutral and negative sentiment detection, and the role of dialect-specific prompting in mitigating common misclassification issues.

5.1 Dialectal Prompting and Models Behaviour

While our results clearly demonstrate that aligning the prompt dialect with the input data leads to improved performance, the extent and nature of this improvement vary notably across models, offering insights into their underlying language handling strategies. DeepSeek, for instance, appears highly sensitive to prompt–input mismatches, performing substantially better when explicitly prompted in the same dialect as the input. This sensitivity suggests a greater reliance on surface-level lexical cues and a narrower contextual understanding of dialectal variation.

By contrast, Qwen2.5 maintained consistently strong performance even in cases of dialect mismatch, particularly when using MSA prompts. This pattern points to a more generalised internal representation of Arabic, enabling the model to adapt across dialects without requiring perfect alignment. LLaMA-3 occupied a middle ground; while its overall performance was slightly lower, it exhibited notable stability across mismatched conditions. This resilience may reflect a more balanced pretraining distribution or stronger contextual abstraction mechanisms.

These contrasts underscore an important methodological takeaway; prompt design must consider not only the linguistic characteristics of the input data, but also the model's capacity for generalisation and sensitivity to variation. Dialectal prompting, therefore, does not yield consistent benefits in all scenarios but a strategy whose effectiveness depends on the interplay between model architecture, training exposure, and the nature of the task.

Model	Dialect	Prompt Before	Prompt After	Neutral F1	Negative F1	Positive F1
Qwen2.5	Darija	MSA	Darija	0.66 (–0.08)	0.78 (+0.04)	0.79 (–0.01)
	Saudi	MSA	Saudi	0.55 (–0.02)	0.78 (+0.04)	0.79 (–0.02)
	MSA	English	MSA	0.42 (+0.21)	0.30 (+0.00)	0.77 (–0.04)
LLaMA-3	Darija	MSA	Darija	0.00 (+0.60)	0.77 (+0.11)	0.75 (+0.03)
	Saudi	MSA	Saudi	0.55 (–0.12)	0.84 (–0.03)	0.76 (+0.01)
	MSA	English	MSA	0.63 (+0.00)	0.68 (+0.00)	0.73 (+0.00)
DeepSeek	Darija	MSA	Darija	0.37 (+0.21)	0.78 (+0.04)	0.70 (–0.17)
	Saudi	MSA	Saudi	0.13 (+0.42)	0.78 (+0.04)	0.70 (+0.04)
	MSA	English	MSA	0.17 (+0.20)	0.42 (+0.41)	0.73 (+0.00)

Table 5: Class-level F1 improvements after using dialect-specific prompts, shown as after score ±change from before

Example Text	Example in English	True Sentiment	Predicted Sentiment
كتبت على العاج من قبل، وخا مؤخرا كان عندي سبب نرجع نزورهم وفعلاً تصدمت. باين بلي تبدلات اليد العاملة، ودابا كيديرو شي إصلاحات	I've written about Al'aj before, and I recently had a reason to visit again. I was truly shocked. It seems the place has changed staff and is in the process of renovating the place.	Neutral	Positive
الجودة ديالها بالنسبة للثمن، بالنسبة لسكندرية، من أحسن الفنادق بالثمن، والثمن اللي اتخلصو فيه مناسب	The quality is okay to the price. For Alexandria, it is one of the best hotels at the price, the price they get is reasonable.	Neutral	Positive
الغرفة مساحتها زينه بس انها بسيطة ومافيها اي شي من الطابع الافريقي ماعدا اللوحات الموجودة في الغرفة	The room is a good size, but it's simple and doesn't have anything with an African touch except for the paintings in the room.	Neutral	Positive

Table 6: Examples of Model Predictions for the Neutral Class

5.2 Error Patterns and Linguistic Ambiguity

An analysis of misclassified examples reveals persistent challenges in detecting neutral and negative sentiment in Arabic dialectal text. These errors largely stem from ambiguity in tone, indirect phrasing, and the models' overreliance on surface-level sentiment cues.

Examples of neutral sentiment misclassifications are shown in Table 6, highlighting how LLMs can struggle to distinguish between implicit evaluation and actual sentiment, particularly when affect is downplayed or absent. Models frequently defaulted to positive predictions when sentences included general approval or value-for-money statements, even when the intent was descriptive or comparative. In such cases, expressions of adequacy or appropriateness were misread as praise.

Negative sentences were often mislabeled as neutral due to polite tone, narrative framing, or indirect complaint structures. As shown in Table 7, this included critiques masked in formal language, sarcastic remarks, or subtle indicators of dissatisfaction, pointing to a core limitation in generative models when dealing with understatement or pragmatically encoded negativity, both common in Arabic discourse.

Example Text	Examples in English	True Sentiment	Predicted Sentiment
مخيب للآمال ممكن يكون نجمة وحده حجزت عن طريق فندق نزلنا فيه في قرطبة	Disappointing, it might as well be a one-star hotel. I booked through a hotel we stayed at in Córdoba	Negative	Positive
رغم ان الفندق كان غالي الا انى اخترت انى احجز فيه عشان شهرته وتقييمات النزلاء له	Even though the hotel was expensive, I chose to book it because of its reputation and guest reviews	Negative	Neutral
بوفيه الفطور يحتاج تتنوع اكثر الإدارة مما تهتم بالملاحظات وكأنك مقيم مجانا.	The breakfast buffet needs more variety. Managers don't care about feedback, as if you're staying there for free	Negative	Positive
كاين الصداع بزاف النادي الي قريب كيخدم الليل كامل وكتسمع حتى الطيارات كتقلع وكتهبط	There's a lot of noise; the nearby club operates all night and you can even hear planes taking off and landing	Negative	Neutral

Table 7: Examples of Model Predictions for the Negative Class

The use of dialect-specific prompts and alignment between prompt and input varieties contributed to reducing such errors. When instructions matched the input dialect, models demonstrated improved handling of tone, idiomatic phrasing, and pragmatic nuance. This alignment reduced the overprediction of the positive class and supported more

accurate recognition of subtle sentiment in dialectal Arabic.

6 Conclusion and Future Work

This study presented a comparative evaluation of three large language models, DeepSeek-R1, Qwen2.5, and LLaMA-3, for sentiment classification across Modern Standard Arabic (MSA), Saudi Arabic, and Darija. Through a combination of dialect-specific prompting, we investigated how model behaviour varies across dialects and prompt configurations. Our results show that alignment between input and prompt significantly improves performance.

Qwen2.5 achieved the highest overall accuracy and F1 scores, demonstrating strong generalization even under prompt mismatch. DeepSeek, while more sensitive to alignment, showed substantial gains when prompted in the target dialect. LLaMA-3 maintained consistent performance, balancing robustness and sensitivity. Across all models, neutral and negative sentiment classification remained the most challenging, often due to indirect phrasing, implicit affect, or sarcasm.

The findings highlight how dialectal prompting benefits underrepresented varieties and improves recognition of neutral and negative sentiment, which remain difficult due to pragmatic ambiguity and indirect affective cues. Performance gains were especially evident in DeepSeek, while Qwen2.5 showed more robust generalisation. LLaMA-3 performed moderately but consistently, reflecting a trade-off between flexibility and dialectal sensitivity. In all cases, dialect-aware prompting reduced overprediction of the positive class and supported more accurate sentiment interpretation.

Future directions include expanding to additional Arabic varieties. We also plan to evaluate zero-shot and few-shot prompting strategies. Finally, extending the task to include aspect-based sentiment analysis would provide a more fine-grained understanding of how models handle sentiment in specific topics or entities within complex dialectal texts.

References

Ahmed Suliman Abuhammad and Mahmoud Ali Ahmed. 2024. Automatic negation detection for semantic analysis in arabic hotel reviews through lexical and structural features: A supervised classification. *Journal of Information and Communication Technology*, 23:709–744.

Mohammad Al-Smadi, Baraa Talafha, Mahmoud Al-Ayyoub, and Yaser Jararweh. 2019. Using long short-term memory deep neural networks for aspect-based sentiment analysis of arabic reviews. *International Journal of Machine Learning and Cybernetics*, 10(8):2163–2175.

Abdulmohsen Al-Thubaity, Sakhar Alkhereyf, Hanan Murayshid, Nouf Alshalawi, Maha Bin Omirah, Raghad Alateeq, Rawabi Almutairi, Razan Alsuwailem, Manal Alhassoun, and Imaan Alkhanen. 2023. Evaluating chatgpt and bard ai on arabic sentiment analysis. Technical report. In this study they use Saudi tweets corbus for positive negative and netral SA. They apply GPT3.5 and GPT also PaLM for the task.

Alanoud Alotaibi and Farrukh Nadeem. 2024. Leveraging social media and deep learning for sentiment analysis for smart governance: A case study of public reactions to educational reforms in saudi arabia. *Computers*, 13(11).

Asma Ameur, Sana Hamdi, and Sadok Ben Yahia. 2023. Sentiment analysis for hotel reviews: A systematic literature review. *ACM Comput. Surv.*, 56(2).

T. Suresh Balakrishnan, P. Gururama Senthilvel, U. Samson Ebenezar, L. Karthikeyan, and B. S. Kishan. 2025. Exploring few-shot learning to enhance nlp's cross-domain capabilities. In *Proceedings of 2025 International Conference on Computing for Sustainability and Intelligent Future, COMP-SIF 2025*. Institute of Electrical and Electronics Engineers Inc.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Salma Khaled, Ensaf Hussein Mohamed, and Walaa Medhat. 2024. Evaluating large language models for arabic sentiment analysis: A comparative study using retrieval-augmented generation. In *Procedia Computer Science*, pages 363–370.

Jan Ole Krugmann and Jochen Hartmann. 2024. Sentiment analysis in the age of generative ai. *Customer Needs and Solutions*, 11(1):3.

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

Ahmad S. Mohammad, Omar Qwasmeh, Baraa Talafha, Mahmoud Al-Ayyoub, Yaser Jararweh, and Elhadj Benkhelifa. 2016. An enhanced framework for aspect-based sentiment analysis of hotels' reviews: Arabic reviews case study. In *2016 11th International Conference for Internet Technology and Secured Transactions (ICITST)*, pages 98–103. IEEE.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Ahmad S. Mohammad, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30.

Faisal Qarah and Tawfeeq Alsanoosy. 2025. Evaluation of arabic large language models on moroccan dialect. *Technology & Applied Science Research*, 15(1):22478–22485.

Petr Rossyaykin. 2024. Structured sentiment analysis using few-shot prompting of an ensemble of llms. This paper is a shared task paper. The outhor trains the LLMs (Gork-Beta, GPT-4o, DeepSeek and Menstral Large 2) on different satages based on prompts. Stage 1 is for the basic prompts. The second is for Augmentd prompts. They Fine-tune the models and used 12 examples per prompt. They then added A THIRD STAGE for text embedding generations.

Sameh M. Sherif, A. H. Alamoodi, O. S. Albahri, Salem Garfan, A. S. Albahri, Muhammet Deveci, Mohammed Rashad Baker, and Gang Kou. 2023. Lexicon annotation in sentiment analysis for dialectal arabic: Systematic review of current trends and future directions (article). *Information Processing and Management*, 60.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

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.

Mohamed Zouidine and Mohammed Khalil. 2025. Large language models for arabic sentiment analysis and machine translation. *Engineering, Technology and Applied Science Research*, 15:20737–20742.