

Sentiment Swapping in Dialectal Arabic: Dataset Construction and Model Performance Analysis

Osamah E. Alnahari and Shadi I. Abudalfa*

Cite <https://doi.org/10.64589/juri/231539>

Submitted: November 29, 2025 Revised: April 20, 2026 Accepted: April 28, 2026

ABSTRACT

In this study, we examine sentiment swapping in Arabic dialects utilizing large language models (LLMs), which is a task that requires the reversal of sentiment polarity while preserving meaning and dialect. Five Arabic sentiment datasets were filtered and validated to construct a high-quality dialectal corpus, and multiple 7B to 13B models were evaluated under zero- and five-shot prompts. The results indicate that few-shot examples consistently improve performance across all models. Qwen2.5-7B-Instruct achieved the strongest sentiment inversion, whereas ALLaM-7B-Instruct-preview demonstrated the best dialect preservation and semantic similarity. Overall performance remained moderate, highlighting the difficulty of polarity-controlled rewriting in dialectal Arabic. This study provides a benchmark for future Arabic generative natural language processing research, as well as a carefully selected sentiment-swapped dataset.

Keywords: Arabic Dialect, Sentiment Swapping, Large Language Model, Natural Language Processing, Polarity Inversion, Dataset Construction

1. INTRODUCTION

Sentiment-aware language technologies are becoming increasingly important in domains such as marketing, politics, and customer engagement based on the rapid development of digital communication. Arabic sentiment classification remains a significant challenge, even though natural language processing (NLP) models have achieved impressive results in English sentiment classification. Arabic's rich morphology, diverse dialects, and small number of labeled datasets have led to relatively slow and dispersed progress. Only a single aspect of sentiment analysis has been the focus of most previous studies, namely, simple classification tasks.

A more ambitious task is sentiment swapping, which involves rewriting text to express the opposite sentiment, while preserving the original meaning. Sentiment swapping can expand training data, improve dialogue systems, and strengthen content moderation. However, sentiment swapping remains largely unexplored in the Arabic language. With the assistance of large language models (LLMs), creating parallel sentiment-swapped corpora and design models for controlled rewriting is now feasible. This study aims to fill an important research gap by focusing on Arabic dialects and offering new resources and methodologies that can enrich low-resource NLP and broaden the scope of sentiment-based applications.

The structure of the remaining paper is as follows. Section 2 is a literature review. Section 3 describes the methodologies. Section 4 explains the experimental setup. Section 5 provides the results and discussions. Section 6 concludes the study.

2. LITERATURE REVIEW

Research on Arabic sentiment processing has primarily focused on classification, while scant attention has been devoted to sentiment swapping. Early work revealed that general neural machine translation models often fail to preserve sentiment when translating Arabic reviews, sometimes even reversing polarity, although fine-tuning significantly improves preservation¹. More recent evaluations of ChatGPT and Bard on Saudi dialect tweets found GPT-4 to be competitive with Arabic-BERT models in terms of classification performance but relatively weak in terms of generating high-quality dialect text with altered sentiment². Together, these studies highlight both the potential and limitations of current models for handling polarity shifts in Arabic.

Recent studies have mainly focused on LLMs built specifically for Arabic dialects. Atlas-Chat, which is a family of models fine-tuned for Moroccan Darija, demonstrates clear gains in dialect identification and generation, highlighting how targeted pre-training can boost accuracy under low-resource conditions³. AceGPT is an Arabic-aligned LLM designed based on regional linguistic norms that has outperformed general models on several Arabic benchmarks, confirming that localized alignment can significantly improve both generation and understanding⁴. Taken together, these efforts demonstrate why sentiment-swapping systems must account for dialect differences, rather than depending solely on broad multilingual models.

Advances in sentiment style transfer beyond Arabic have also highlighted the importance of controlled polarity inversion. One foundational framework introduced sentiment swapping

by removing and regenerating sentiment-bearing expressions without requiring parallel data⁵. Another study stressed that polarity flipping is central to sentiment transfer and proposed polarity-aware denoising to reduce meaning loss during rewriting⁶. Within Arabic, domain-specific formality-transfer systems have demonstrated that Arabic-oriented LLMs preserve meaning and linguistic identity better than general models⁷.

Complementary research has explored more direct manipulations of sentiment. For example, adversarial synonym substitution has been shown to flip classifier predictions reliably while maintaining naturalness in Arabic text⁸. Additionally, models such as ArSarcasm capture implicit sentiment inversion through sarcasm⁹. Despite this progress, the scarcity of dialect-specific datasets continues to limit the development of robust sentiment-swapping models for Arabic dialects.

3. METHODOLOGY

3.1. Approach. Our methodology follows a structured multistage pipeline designed to construct a high-quality Arabic sentiment-swapped dataset. As illustrated in Figure 1, the process began with the collection and filtering of multiple dialectal sentiment datasets to derive clean, polarity-defined inputs that are suitable for controlled rewriting. Several LLMs were evaluated using a representative sample to identify the model most capable of producing accurate sentiment inversions while preserving meaning and dialect. After selecting the best-performing model, the full dataset underwent sentiment swapping, followed by automated verification steps utilizing computer-aided mechanistic laboratory (CAMEL) to check whether sentiments were actually swapped and multidialectal Arabic bidirectional encoder representations from transformers (MarBERT)v2 to ensure that the dialect was preserved. Only sentences that passed all quality filters were retained, resulting in a reliable and dialect-preserving sentiment-swapped corpus.

Recent studies on prompting-based style transfers have supported the design choices adopted in this pipeline. LLMs have been shown to perform arbitrary text style transfer using natural-language prompts, demonstrating that zero-shot or few-shot prompting can replace task-specific fine-tuning in many rewriting tasks¹⁰. A complementary study introduced the prompt-and-rerank method, where models generate multiple candidate outputs, and a downstream scoring mechanism selects the best output based on content preservation and stylistic accuracy¹¹. This approach is closely aligned with our verification strategy, where sentiment and dialect classifiers (CAMEL and MarBERTv2) serve as evaluators to ensure that the generated sentences invert polarity while maintaining the original dialect. These methodological foundations validate the use of prompting-based sentiment rewriting in our study.

3.2. Dataset. Our dataset construction process was designed to produce a robust and dialect-focused corpus for Arabic sentiment swapping. Five publicly available Arabic sentiment datasets were used. Collectively, these datasets cover a wide range of dialectal variations across the Arab world, containing over 100,000 samples before filtering. The primary characteristics of the datasets are summarized in Table 1.

The preprocessing and filtering steps were carefully defined to ensure that the final dataset contained only dialectal Arabic sentences with clear sentiment polarity. The first stage involved removing all samples labeled as neutral and retaining only those classified as positive or negative. This restriction simplified the sentiment polarity dimension and ensured that all entries could be utilized effectively during the sentiment swapping process, where distinct polarity differences are necessary.

Following sentiment selection, dialect classification was applied to identify and exclude samples written in modern standard Arabic (MSA). Because this study focused on Arabic dialects, MSA samples were considered out of scope. The classification was performed using the MarBERTv2 Arabic written

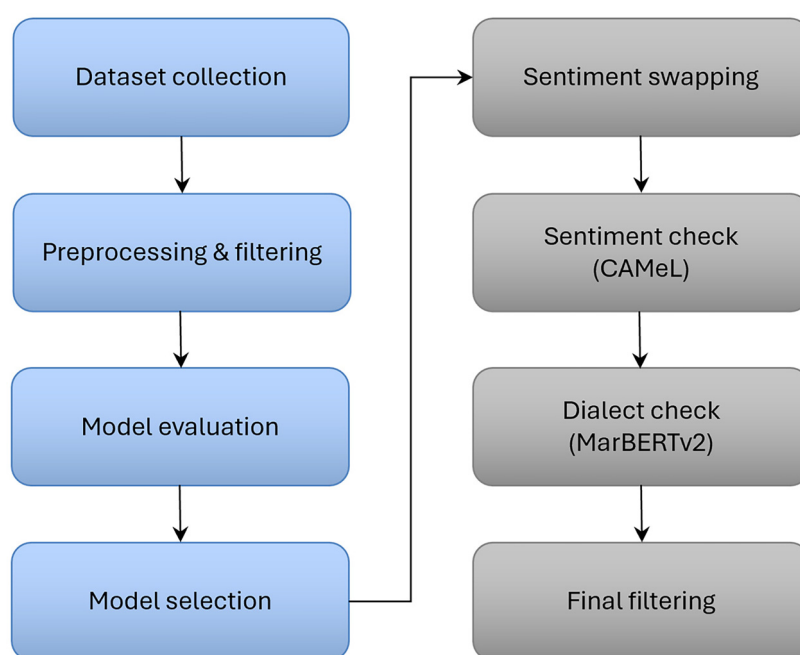


Figure 1. Overall workflow for constructing the Arabic sentiment-swapped dataset

Table 1. Summary of collected Arabic sentiment datasets

Title	Dialect	Samples	Final selected samples	Type	Source
Algerian Arabic Sentiment Analysis Dataset	Algerian	1,500	292	Comments	Kaggle ³
Arabic Egypt Customer Reviews	Egyptian	40,046	4,960	Reviews	Kaggle ⁴
Arabic Sentiment Twitter Corpus	Multiple	58,700	7,394	Tweets	Kaggle ⁵
Twitter Dataset for Arabic Sentiment Analysis	Multiple	2,000	208	Twitter	Archive ⁶
Sentiment Analysis for Moroccan Dialect	Moroccan	4,337	120	Comments	GitHub ⁷

³Algerian Arabic Sentiment Analysis Dataset⁴Arabic Egypt Customer Reviews⁵Arabic Sentiment Twitter Corpus⁶Twitter Dataset for Arabic Sentiment Analysis⁷Sentiment Analysis for Moroccan Dialect

dialect classifier, which assigns each text a predicted dialect label with a confidence score. Only non-MSA samples with a confidence score greater than or equal to 0.9 were retained. This confidence threshold was selected to ensure high-quality dialectal identification and minimize noise from ambiguous cases.

Although the general selection procedure was consistent across all datasets, a few datasets required specific additional refinements. For example, in the Arabic Egypt Customer Reviews dataset, very short sentences containing fewer than five words or 20 characters were excluded. Such short entries often lack

sufficient sentiment information and are too similar to MSA, which could have reduced dialectal clarity. In the Arabic Sentiment Twitter Corpus, only samples with a sentiment label confidence score of one were retained. These values were provided within the metadata of the dataset, and we ensured that only highly reliable sentiment annotations were included.

After applying these criteria, the resulting corpus represented a balanced and high-quality dataset of Arabic dialectal sentences labeled according to their sentiment polarity. This refined dataset served as the foundation for subsequent experiments on Arabic

Table 2. Models and prompts used for sentiment swapping evaluation

Model	Prompt
DeepSeek-R1-Distill-Llama-70B	Sentence (in {dialect}Arabic dialect): "{sentence}" Current sentiment: {sentiment} Rewrite the sentence with the opposite sentiment. Preserve the meaning and context, keep the same dialect. Output ONLY JSON with this format: {"swapped_sentence": "<your sentence>"}. Do NOT include any explanation, reasoning, or extra text.
DeepSeek-R1-Distill-Qwen-32B	Sentence (in {dialect}Arabic dialect): "{sentence}" Current sentiment: {sentiment} Rewrite the sentence with the opposite sentiment. Preserve the meaning and context, only change the sentiment. The swapped sentence should be in arabic only Output only in JSON: {"swapped_sentence": "<your sentence>"}
Llama-3.3-70B-Instruct	Sentence (in {dialect}Arabic dialect): "{sentence}" Current sentiment: {sentiment} Rewrite the sentence with the opposite sentiment. Preserve the meaning and context, only change the sentiment. The swapped sentence should be in Arabic only. Output only in JSON: {"swapped_sentence": "<your sentence>"}
Jais-Adapted-70B-Chat	Sentence (in {dialect}Arabic dialect): "{sentence}" Current sentiment: {sentiment} Rewrite the sentence with the opposite sentiment. Preserve the meaning and context, only change the sentiment. The swapped sentence should be in Arabic only Output ONLY the rewritten sentence without explanation. swapped sentence:
Jais-Family-30B-16k-Chat	Sentence (in {dialect}Arabic dialect): "{sentence}" Current sentiment: {sentiment} Rewrite the sentence with the opposite sentiment: Preserve the meaning and context, only change the sentiment. Output only the rewritten sentence. ### Response:

Table 3. Performance comparisons of five models on sentiment-swapping evaluation using 200 random samples (100 positive and 100 negative)

Model	Opposite swaps	Neutral	No change	Avg. cosine similarity
Llama-3.3-70B-Instruct	117	33	50	0.9371
JAIS-Family-30B-16K-Chat	91	34	75	0.8971
JAIS-Adapted-70B-Chat	85	56	59	0.8045
DeepSeek-R1-Distill-Qwen-32B	86	39	77	0.9545
DeepSeek-R1-Distill-Llama-70B	128	34	38	0.8937

sentiment swapping, ensuring that both sentiment and dialect were consistently preserved throughout the data generation and evaluation processes.

We randomly selected 200 samples (100 positive and 100 negative) from the filtered dataset to evaluate the capability of each model to swap the sentiments of dialectal Arabic sentences while preserving their dialectal characteristics. Five LLMs with parameter counts ranging from 30B to 70B were selected because these models are typically more reliable for handling complex Arabic dialects and are sufficiently large to generate high-quality text. Early tests revealed that each model responded differently to wording and formatting changes, so the prompts for each model were modified accordingly. The final prompt versions used for each model are summarized in Table 2. These prompts were chosen to ensure stable outputs and allow for fair and reliable comparisons of their sentiment-swapping performance.

The generated sentences were classified using the computational approaches to modeling language laboratory bidirectional encoder representations from transformers base architecture Arabic language computational approaches to modeling language BERT dialectal Arabic sentiment analysis (CAMEL-Lab/bert-base-arabic-Camelbert-da-sentiment) model, which assigns each sentence a positive, negative, or neutral label. This step allowed us to measure how effectively each model reversed the sentiments of the input sentences while preserving their dialect features. The

evaluation results, including the number of unchanged sentences, number of successful sentiment swaps, swap rate percentage, and average similarity between original and swapped sentences, are summarized in Table 3.

As summarized in Table 3, the DeepSeek-R1-Distill-Llama-70B model produces the highest number of successful sentiment flips with 128 opposite swaps, outperforming all other models in this evaluation. Although Llama-3.3-70B-Instruct achieves a slightly higher similarity score (0.937), it completes fewer opposite swaps at 117. DeepSeek-R1-Distill-Llama-70B delivers the strongest balance, effectively reversing sentiment while preserving meaning, coherence, and dialectal authenticity, with an average similarity of 0.8937. By contrast, JAIS-Family-30B-16K-Chat and DeepSeek-R1-Distill-Qwen-32B exhibit less consistent behavior, achieving only 91 and 86 opposite swaps, respectively. When considering all metrics, DeepSeek-R1-Distill-Llama-70B stands out as the most reliable option and was selected as the main model for generating the final sentiment-swapped dataset.

To ensure that the selected model could reliably swap sentiments, we first verified whether it understood the original sentiments represented by every sample in the dataset. As summarized in Table 4, the DeepSeek-R1-Distill-Llama-70B model was given a sentiment classification prompt that asked it to label each sentence as positive, negative, or neutral. We then compared its predictions with the ground-truth labels and retained only sentences

Table 4. Prompts used for verification and sentiment-swapping tasks when using the DeepSeek-R1-Distill-Llama-70B model

Task	Prompt used	Purpose
Sentiment classification	You are an expert Arabic sentiment classifier that understands dialectal Arabic. Classify the following sentence into one of: Positive, Negative, or Neutral. Examples: 1. Sentence: "أنا مبسوط اليوم", sentiment: Positive 2. Sentence: "هالشى مو زين أبد", sentiment: Negative 3. Sentence: "الوضع عادي ما تغيّر شي", sentiment: Neutral Output ONLY JSON in this exact format: {"sentiment": "<Positive Negative Neutral>"} Sentence: "{sentence}"	Verify that the model understands the original sentiment of all samples
Sentiment swapping	Sentence (in {dialect}Arabic dialect): "{sentence}" Current sentiment: {sentiment} Rewrite the sentence with the opposite sentiment. Preserve the meaning and context, keep the same dialect. Output ONLY JSON with this format: {"swapped_sentence": "<your sentence>"} Do NOT include any explanation, reasoning, or extra text.	Generate sentences with the opposite sentiment while preserving dialect

that were classified correctly. This step confirmed that the model had a solid grasp of each input before attempting to flip the sentiments.

After filtering, we used the sentiment swapping prompt from Table 4 to perform the actual inversion. The model was asked to rewrite every sentence with the opposite sentiment while maintaining the original meaning and dialectal features. A generation limit of 1,000 tokens per sample was set to prevent truncation, and incomplete outputs were removed. This step ensured that all remaining swapped sentences were complete, coherent, and suitable for evaluating both polarity reversal and dialect preservation. As a final quality-control step, all swapped sentences were reevaluated for both dialect preservation and correct polarity inversion. Each generated sentence was first passed through the same MarBERTv2 dialect classifier used during dataset filtering, and any sample whose predicted dialect did not match the corresponding original dialect was removed to ensure that rewriting did not drift toward MSA or another dialect. The sentences were then reclassified using the CAMEL sentiment model employed previously, and only those whose swapped versions received the opposite sentiment label were retained. After applying these two verification filters for dialect consistency and accurate sentiment reversal, the resulting dataset contained 12974 high-quality sentiment-swapped samples.

4. EXPERIMENTAL SETUP

All experiments were conducted using the verified dialectal dataset described in the previous section, which consisted of 12974 samples. To ensure controlled comparisons, each model was evaluated on the same dataset under identical prompting, generation, and scoring conditions. Each model was configured with a maximum generation limit of 2048 tokens to ensure sufficient space for internal reasoning while guaranteeing that the final output contained a completely swapped sentence.

Two prompting configurations were tested: zero-shot and five-shot. These settings were selected to evaluate the extent to which the models benefited from explicit guidance when performing sentiment swapping in Arabic. The zero-shot configuration tested the inherent ability of a model to understand a task without examples, offering a baseline for general reasoning and instruction-following capabilities. By contrast, the five-shot

configuration provided a model with several illustrative examples of sentiment inversion, which was particularly valuable for tasks involving subtle polarity changes and dialect preservation. Previous studies on controlled text generation have demonstrated that few-shot prompts often improve performance by clarifying the transformation patterns expected from the model, particularly in low-resource or dialect-specific scenarios. All models shared the same unified prompt template, and the examples used in the five-shot setting were fixed across experiments. These examples were drawn from external dialectal sentences that were not included in the target dataset to avoid data leakage while ensuring consistent comparisons across all models.

4.1. Models Evaluated. Six model configurations were included in our experiments, covering three Arabic-capable LLMs evaluated under zero- and five-shot prompting settings. These configurations were intentionally limited to lightweight models in the 7B to 13B parameter range to examine whether smaller models could reliably perform controlled sentiment rewriting in dialectal Arabic without relying on extremely large or resource-intensive systems. Additionally, only the instruct/chat variants of the models were selected because these models were optimized for following instructions and generating structured text, making them more suitable for sentiment-swapping tasks compared with the baseline or pretrained models. Figure 2 illustrates the final set of evaluated models and corresponding prompting configurations.

Beyond model selection, our experimental setup required the design of a consistent prompting strategy for sentiment swapping. Table 5 summarizes the full few-shot prompt template utilized in the five-shot configuration, along with the examples of sentence pairs supplied to the models. These examples illustrated the exact transformation expected from the model, flipping the sentiment while preserving the original meaning and dialect. This template was fixed across all experiments to ensure fair comparisons. Including these examples was particularly important for models that struggled with instruction alignment, as the examples provided explicit guidance regarding the type of rewriting required for the target task.

4.1.1. Evaluation metrics. To assess the performance of each model across different prompting configurations, a set of quantitative metrics was used to measure sentiment correctness,

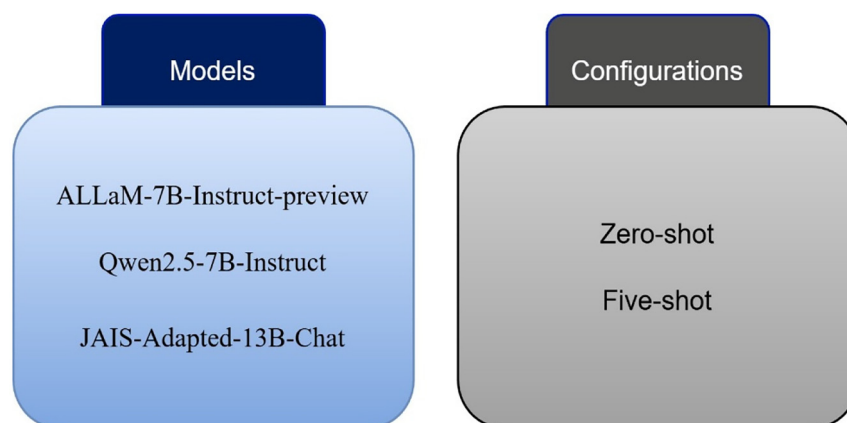


Figure 2. Overall framework used for benchmarking on the Arabic sentiment-swapped dataset

Table 5. Few-shot prompt and example pairs used for sentiment swapping

Prompt used	Few-shot samples
{FEW_SHOT_EXAMPLES}	sentence: "المطعم ده تعبان قوي والخدمة بطيئة بشكل مزعج" swapped_sentence: "المطعم ده ممتاز والخدمة سريعة ومريحة" 2) sentence: "الأكل كان كتير مقرف وما إلو أي معنى" swapped_sentence: "الأكل كان كتير ممتع وفيه معنى حلو" 3) sentence: "هالمنتج رديء ولا يستحق الفلوس أبداً" swapped_sentence: "هالمنتج ممتاز ويستاهل كل فلس" 4) sentence: "الخدمة فالشركة كارثية وما يعاونوك فوالو" swapped_sentence: "الخدمة فالشركة زويينة وكيعاونوك فكلشي" 5) sentence: "السيارة مزعجة كلش وصوتها يضوج" swapped_sentence: "السيارة هادئة كلش وصوتها ما ينسمع"
NOW YOUR TURN:	
Sentence (in {dialect}Arabic dialect): {text}	
Current sentiment: {label}	
Rewrite the sentence with the opposite sentiment.	
Preserve the meaning and context, only change the sentiment	
The swapped sentence should be in Arabic only.	
Output ONLY in JSON: {{\ "swapped_sentence\": \"<your sentence>\\"}}	

semantic preservation, and dialect consistency. These metrics facilitated the comprehensive evaluation of the models' ability to invert sentiment while maintaining both meaning and dialectal features. The full list of the metrics is summarized in Table 6. These metrics captured both the correctness of the sentiment inversion and the linguistic fidelity of the rewritten sentences, supporting the well-rounded evaluation of model performance.

5. RESULTS AND DISCUSSION

5.1. Results. Table 7 summarizes the performance of all six model configurations across four of the evaluation metrics. Overall, the results reveal that five-shot prompting consistently improved performance across all models. Qwen2.5-7B-Instruct achieves the highest sentiment inversion quality, with its five-shot configuration reaching the highest F1-score of 0.74, surpassing the value of 0.70 in the zero-shot setting. This pattern clearly indicates that providing examples helped the model gain a better understanding of the sentiment-swapping task.

In terms of similarity to the ground-truth swapped sentences, ALLaM-7B-Instruct-preview achieves the best overall alignment, with its five-shot configuration yielding the lowest TER (49.28), highest BLEU score (38.73), and highest BERTScore (0.871), suggesting that the generated swaps were the most similar to

the reference sentences in terms of both lexical and semantic similarity.

With an accuracy of 0.248 and a BLEU score of 16.13 in the zero-shot configuration, JAIS-adapted-13B-chat performs the worst on the majority of metrics. However, the five-shot prompt significantly improves the results, increasing the dialect accuracy from 0.561 to 0.789, accuracy to 0.376, and BLEU to 28.72. JAIS benefits the most from few-shot prompting in terms of absolute gains, despite lagging behind the other models. Overall, the results indicate that Qwen2.5-7B-Instruct performs best in terms of sentiment inversion, ALLaM-7B-Instruct-preview performs best in terms of preserving meaning and dialect, and JAIS exhibits weaker baseline performance with the largest improvements when few-shot examples are provided.

Overall, the results demonstrate that the model performance is generally weak to moderate, rather than strong, on this task. The F1-scores, which range from 0.329 to 0.74, reveal that even the best-performing model struggles to flip sentiments consistently. Similarly, BLEU scores are relatively low, usually ranging from 20 to 38, indicating significant lexical changes and frequent deviations from the expected swapped sentences. Although potentially acceptable, BERTScore values between 0.715 and 0.871 are still below the ranges typically linked with excellently controlled rewriting. Given the difficulty of the target task, where sentiment swapping in Arabic dialects necessitates

Table 6. Evaluation metrics used in our experiments

Metric	Description
Sentiment accuracy	Percentage of sentences where the swapped sentence output received the correct opposite sentiment label.
Sentiment F1-score	Harmonic mean of precision and recall for the sentiment classification of swapped sentences.
Sentiment Bilingual Evaluation Understudy (BLEU) ¹²	Measures lexical similarity between the generated swapped sentence and ground-truth swapped reference.
Sentiment BERTScore ¹³	Measures semantic similarity between the generated swapped sentence and ground-truth swapped reference.
Sentiment Translation Edit Rate (TER) ¹⁴	Edit distance between the generated swapped sentence and ground-truth reference (lower is better).
Dialect accuracy	Measures whether the original dialect was preserved after rewriting.

Table 7. *Sentiment-swapping results*

Model	F1-score	BLEU	BERTScore	TER
ALLaM-7B-Instruct-preview	0.59	37.75	0.869	49.97
ALLaM-7B-Instruct-preview – 5 shots	0.60	38.73	0.871	49.28
Qwen2.5-7B-Instruct	0.7	24.94	0.816	68.79
Qwen2.5-7B-Instruct – 5 shots	0.74	26.46	0.823	65.16
jais-adapted-13b-chat	0.329	16.13	0.715	78.09
jais-adapted-13b-chat – 5 shots	0.41	28.72	0.832	60.41

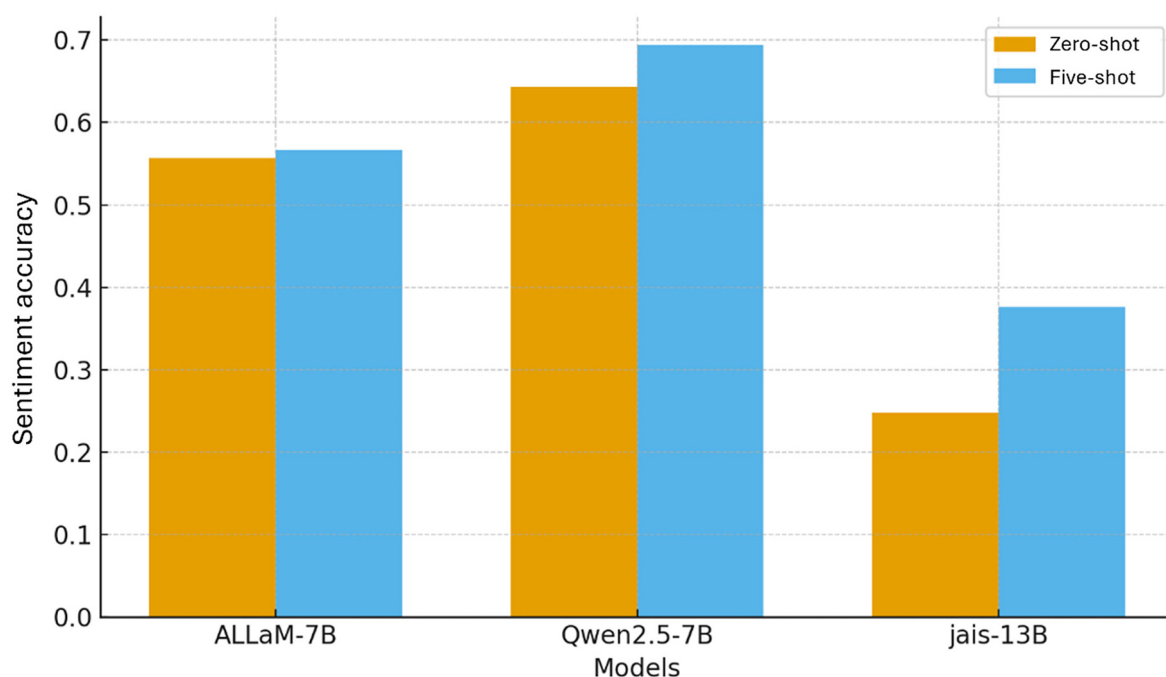
accurate polarity inversion, consistent meaning preservation, and dialect stability without large parallel training corpora, these limitations are to be expected. The relatively high TER values further reinforce the idea that the models often rewrote sentences heavily, rather than performing minimal targeted edits.

Overall, our findings indicate that current models, even those in the 7B–13B range, are not yet sufficiently optimized for fine-grained sentiment manipulation in dialectal Arabic, highlighting the difficulty of the task and the requirement for more specialized alignment or training techniques. The sentiment accuracy results in Figure 3 demonstrate that the overall model performance was mediocre. With the highest accuracy, Qwen2.5-7B-Instruct reaches a value of 0.643 in the zero-shot setting, which improves to 0.694 with five-shot prompting, indicating that this model is relatively consistent at flipping sentiments. With accuracies of 0.557 and 0.567, ALLaM-7B-Instruct-preview performs similarly but at a lower level, exhibiting only slight improvements with the addition of few-shot examples. The JAIS-adapted-13B-Chat model performs poorly overall, with a very low zero-shot accuracy of 0.248, while five-shot prompting increases its accuracy to 0.376. With only Qwen2.5 exhibiting consistently strong behavior and the other models providing limited or insufficient accuracy for this task, these results demonstrate the difficulty of

sentiment inversion in Arabic dialects, as well as the different degrees of instruction alignment across models.

The dialect accuracy results presented in Figure 4 reveal noticeable gaps between the models in terms of their ability to preserve dialectal characteristics during sentiment swapping. ALLaM-7B-Instruct-preview performs the best, achieving consistently high values of 0.90 in the zero-shot setting and 0.907 with few-shot prompting, indicating that its rewrites remained highly faithful to the original dialect. Qwen2.5-7B-Instruct exhibits moderate dialect preservation, improving from 0.669 to 0.715 with few-shot examples but still lagging behind ALLaM. By contrast, the JAIS-adapted-13B-Chat model struggles significantly, with a low zero-shot accuracy of 0.561, suggesting a frequent drift toward MSA or unintended dialect shifts. However, its five-shot accuracy increases substantially to 0.789, demonstrating that this model benefits significantly from explicit guidance. Overall, these results indicate that although ALLaM reliably maintains the original dialect, the other models exhibit weaker and sometimes unstable dialect consistency, underscoring the difficulty of preserving dialect-specific linguistic features during controlled rewriting.

The results indicate that the objectives of this study were achieved, as the proposed methodology enabled the creation of a

**Figure 3.** Sentiment accuracy results

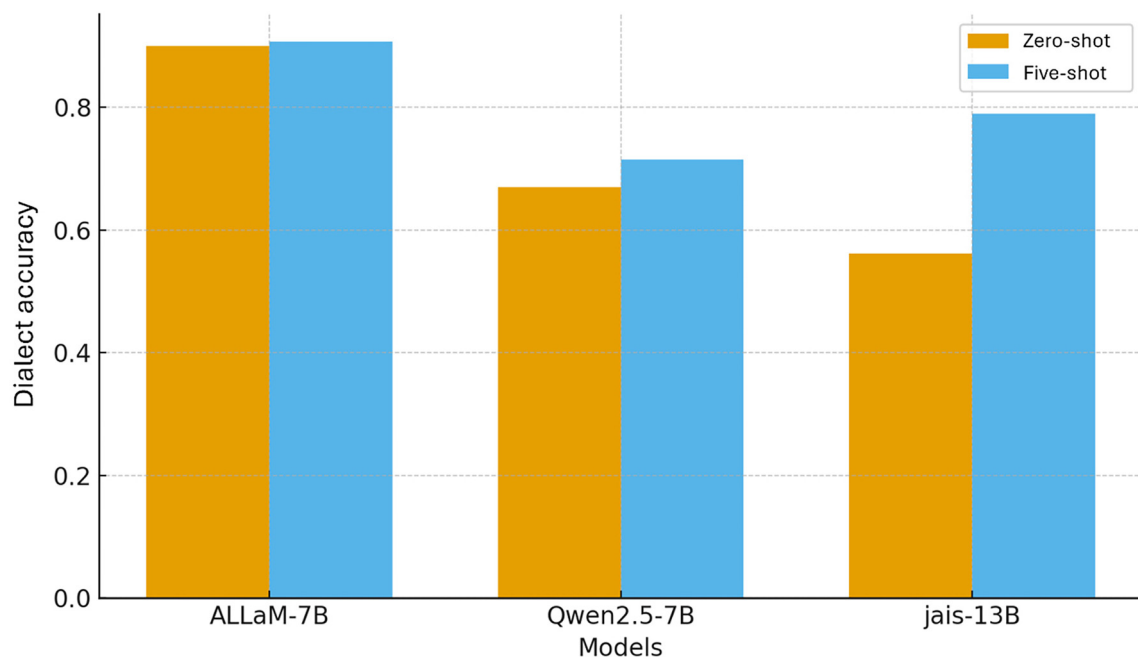


Figure 4. Dialect accuracy results

validated sentiment-swapped dataset for Arabic dialects and the evaluation of multiple large language models on the sentiment-swapping task.

The findings support the hypothesis that large language models can perform sentiment inversion in dialectal Arabic, although performance remains moderate and varies across models. These results are consistent with previous studies on controlled text generation and style transfer, while highlighting the importance of dialect-specific adaptation in Arabic NLP. However, they differ from studies reporting stronger performance in high-resource languages, reflecting the challenges of Arabic dialect diversity and limited resources. Limitations include the reliance on automated evaluation metrics, the restricted set of models, and the coverage of only a subset of Arabic dialects. Future research can address these limitations by improving model robustness, dialect coverage, and sentiment control accuracy.

6. CONCLUSIONS

This paper presented comprehensive evaluations of sentiment swapping in Arabic dialects using a curated dataset and multiple LLMs. The results show that although all models were capable of performing polarity inversion, their overall performance remained moderate across sentiment accuracy, F1-score, BLEU, BERTScore, TER, and dialect preservation metrics. Few-shot prompting consistently improved performance, with Qwen2.5-7B-Instruct achieving the strongest sentiment inversion and ALLaM-7B-Instruct-preview demonstrating the highest semantic and dialectal fidelity. The JAIS-based models exhibited the weakest baseline performance but benefited the most from few-shot examples, highlighting the importance of clear task demonstrations when targeting challenging dialectal transformations.

The results also demonstrate that sentiment swapping in dialectal Arabic remains a challenging task, as models often

introduce neutral sentiments, drift toward MSA, or rewrite sentences excessively. This pattern reflects the scarcity of excellent parallel resources and the linguistic complexity of Arabic dialects. Regardless, this study provided a practical benchmark and a thoroughly filtered and validated sentiment-swapped dataset of 12,974 samples, which can aid in future studies on controlled text generation for Arabic.

The findings of this study have important theoretical and practical implications. Theoretically, they contribute to the growing literature on controlled text generation in Arabic by providing empirical evidence on the ability of large language models to perform sentiment swapping while preserving dialectal characteristics. The results also highlight the challenges associated with polarity-controlled rewriting in dialectally diverse and low-resource language settings. Practically, the validated sentiment-swapped dataset developed in this study provides a valuable benchmark for future Arabic NLP research and may support the development, training, and evaluation of sentiment-aware language technologies. Furthermore, the findings offer guidance for researchers and practitioners regarding model selection and prompting strategies for dialectal Arabic text generation tasks.

Future research may expand upon our findings by developing models that more closely match the linguistic features of Arabic dialects. Stability can be enhanced, and meaning drift can be reduced through techniques such as fine-tuning on richer dialectal corpora, improving model alignment for polarity control, and progressively expanding the dataset to cover a wider variety of dialects. Future studies may also investigate additional Arabic dialects, alternative prompting strategies, and domain-specific applications of sentiment swapping. We anticipate that model capacity for controlled sentiment rewriting will improve with increased exposure to sentiment-flipped patterns and dialect-specific structures, facilitating more precise and reliable handling of polarity inversion in real-world Arabic texts.

AFFILIATIONS AND AUTHOR DETAILS

Undergraduate Author

Osamah E. Alnahari – *Department of Information & Computer Science, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia; 0009-0009-3749-6939*
Email: s202183290@kfupm.edu.sa

Corresponding Author

Shadi I. Abudalfa – *Research Scientist, SDAIA-KFUPM JRC for Artificial Intelligence, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia; 0000-0002-2028-144X*
Email: shadi.abudalfa@kfupm.edu.sa

ACKNOWLEDGEMENTS

The authors greatly appreciate the Undergraduate Research Office at King Fahd University of Petroleum & Minerals (KFUPM) for their kind support through an Undergraduate Student Research Grant. The authors also acknowledge the support provided by the Saudi Data and Artificial Intelligence Authority (SDAIA)-KFUPM Joint Research Center for Artificial Intelligence (JRC-AI) at KFUPM.

REFERENCES

- (1) Saadany, H. & Orāsan, C. Is it great or terrible? Preserving sentiment in neural machine translation of Arabic reviews. arXiv 2010.13814 (2020).
- (2) Al-Thubaity, A. et al. Evaluating ChatGPT and Bard AI on Arabic sentiment analysis. In *Proceedings of ArabicNLP 2023*, 335–349 (2023).
- (3) Shang, G. et al. Atlas-Chat: Adapting large language models for low-resource Moroccan Arabic dialect. In *Proc. 1st Workshop on Language Models for Low-Resource Languages (LoResLM)*, 9–30 (2025).
- (4) Huang, H. et al. AceGPT: Localizing large language models in Arabic. In *Proc. NAACL 2024 (Long Papers)*, 8139–8163 (2024).
- (5) Li, J., Jia, R., He, H. & Liang, P. Delete, retrieve, generate: A simple approach to sentiment and style transfer. arXiv 1804.06437 (2018).
- (6) Mukherjee, S., Kasner, Z. & Dušek, O. Balancing the style-content trade-off in sentiment transfer using polarity-aware denoising. In *International Conference on Text, Speech, and Dialogue*, 172–186 (Springer, 2022).
- (7) Abdu, F. J. et al. An empirical evaluation of Arabic text formality transfer: A comparative study. *Lang. Resour. Eval.* 1–61 (2025).
- (8) Alshahrani, N., Alshahrani, S., Wali, E. & Matthews, J. Arabic synonym BERT-based adversarial examples for text classification. arXiv 2402.03477 (2024).
- (9) Farha, I. A. & Magdy, W. From Arabic sentiment analysis to sarcasm detection: The ArSarcasm dataset. In *4th Workshop on Open-Source Arabic Corpora and Processing Tools*, 32–39 (ELRA, 2020).
- (10) Reif, E. et al. A recipe for arbitrary text style transfer with large language models. In *Proc. ACL 2022 (Short Papers)*, 837–848 (2022).
- (11) Suzgun, M. et al. Prompt-and-Rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models. In *Proc. EMNLP 2022*, 2195–2222 (2022).
- (12) Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. BLEU: A method for automatic evaluation of machine translation. In *Proc. 40th Annual Meeting of the Association for Computational Linguistics*, 311–318 (2002).
- (13) Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. BERTScore: Evaluating text generation with BERT. arXiv 1904.09675 (2019).
- (14) Snover, M., Dorr, B. J., Schwartz, R., Micciulla, L. & Makhoul, J. A study of translation edit rate with targeted human annotation. In *Proc. 7th Conf. of the Association for Machine Translation in the Americas (AMTA)*, 223–231 (2006).