MYM at AraSentEval: A Comparative Study of Arabic Sentiment Swap Models

Yumna Hamdy, Mohab ElDamhougy, Yomna Eid, Ensaf Hussien
Nile University, Egypt
Cairo, Egypt
Y.Hamdy2457@nu.edu.eg, M.Tarek2317@nu.edu.eg, YEid@nu.edu.eg, EnMohamed@nu.edu.eg

Abstract

Sentiment swap is a controlled text generation task that rewrites a sentence by inverting its sentiment polarity while preserving semantic content and fluency. In this paper, we present our system for AraSentEval 2026 Subtask 2 on Arabic sentiment swap, a particularly challenging problem due to Arabic's rich morphology and dialectal variation. We investigate multiple modeling paradigms, including encoder–decoder and multilingual approaches, and propose an enhanced system that combines targeted data augmentation and ensemble learning. Specifically, we augment underrepresented dialectal patterns to improve robustness and ensemble two Arabic-focused sequence-to-sequence models, AraBART and AraT5v2. Experiments are conducted on the MA'aks parallel dataset under fine-tuned settings.

Our system ranked first in AraSentEval 2026 Subtask 2, achieving a BLEU score of 43.0, chrF of 65.36, and sentiment preservation accuracy of 0.7554. The results demonstrate that dialect-aware augmentation together with model ensembling substantially improves sentiment-controlled generation in Arabic and establishes strong baselines for future research in low-resource sentiment manipulation.

Keywords: Arabic NLP, sentiment swap, style transfer, AraSentEval, text generation

1. Introduction

Sentiment analysis is traditionally formulated as a classification task that assigns polarity labels to text. Recent advances in neural generation have enabled sentiment swap, where a sentence is rewritten to express the opposite sentiment while preserving meaning and fluency. This task is substantially more challenging than classification because models must accurately identify and modify sentiment-bearing expressions.

Arabic sentiment swap is particularly difficult due to rich morphology, dialectal variation, and frequent implicit sentiment cues. While Arabic sentiment classification has been widely studied, sentiment-controlled generation remains underexplored. The MA'aks dataset and the AraSentEval shared task provide a standardized benchmark for this problem.

In this work, we present our system for AraSentEval 2026 Subtask 2. We combine dialect-aware data augmentation with an ensemble of two Arabic encoder–decoder models, AraBART and AraT5v2. Our approach ranked first in the shared task, demonstrating the effectiveness of augmentation and model ensembling for Arabic sentiment swap.

Contributions: (i) a top-performing AraBART–AraT5v2 ensemble, (ii) dialect-aware augmentation, and (iii) evaluation of Arabic and multilingual models.[1]

[1]The code

2. Related Work

Sentiment style transfer has drawn growing interest in controlled text generation, but Arabic remains under-resourced. Mughaus et al. (Mughaus et al., 2026) introduced MA'AKS, the first manually curated parallel dataset for Arabic sentiment swap, with about 5,000 Modern Standard Arabic sentence pairs of opposite polarity.

Using MA'AKS, they evaluated AceGPT, JAIS, and Llama-3 under zero-shot, few-shot, and fine-tuning settings. Llama-3-8B-Instruct achieved the best zero-shot semantic preservation (BLEU 43.68 PTN / 44.9 NTP), while supervised fine-tuning produced the strongest overall gains, reaching 59.9 PTN / 57.1 NTP (Mughaus et al., 2026).

However, sentiment control remains difficult, especially for dialectal Arabic, where all models show notable performance drops. The AraSentEval shared task (Ezzini et al., 2026) extended evaluation to a larger benchmark, but systematic study of encoder–decoder architectures and full fine-tuning remains limited.

In this work, we address these gaps by evaluating Arabic-specialized encoder–decoder models (AraT5v2 and AraBART), introducing dialect-aware data augmentation, and exploring ensembling to improve sentiment control and semantic preservation.

Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks, pages 269–273

3. Task and Dataset

3.1. Task Definition

Sentiment swap is a controlled generation task that reverses the sentiment of an input sentence while preserving meaning and fluency. Given a sentence x with label $s \in \{positive, negative\}$, the goal is to generate y expressing the opposite sentiment $\bar{s}$. This task is challenging in Arabic due to rich morphology and implicit sentiment cues.

3.2. MA'aks Dataset

We evaluate on the MA'aks dataset, a parallel corpus for Arabic sentiment swap with separate training, validation, and test files covering MSA, Gulf, Levantine, Egyptian, and Moroccan dialects. The training set contains 6,262 pairs (54.58% positive, 45.42% negative), and the validation set includes 1,315 pairs (61.9% positive, 38.1% negative). The test set comprises 647 unlabeled sentences. Each labeled instance pairs a source sentence with a sentiment-inverted rewrite.

4. Methodology

We investigate Arabic sentiment swap using multiple modeling paradigms and propose an ensemble-based system. Our study compares encoder–decoder and decoder-only approaches, analyzes fine-tuning effects, explores dialect-aware augmentation, and introduces a CAMeLBERT-based selection mechanism.

4.1. Arabic Encoder–Decoder Models

We focus on encoder–decoder architectures, which naturally support controlled rewriting. We evaluate two Arabic-focused models: AraT5v2 and AraBART.

We fine-tune AraBART on MA'aks, obtaining substantial gains. AraT5v2 is directly fine-tuned and consistently performs slightly better, confirming the effectiveness of Arabic-pretrained sequence-to-sequence models.

4.2. Multilingual Encoder–Decoder Model

We also evaluate MT5, a multilingual encoder–decoder model. Despite broad language coverage, MT5 performs poorly on Arabic sentiment swap and incurs higher computational cost, indicating that multilingual pretraining alone is insufficient for fine-grained sentiment manipulation.

4.3. Decoder-Only Model

For comparison, we fine-tune the decoder-only GPT-J-6B model. Although it generates fluent text, it struggles to reliably invert sentiment and consistently underperforms relative to encoder–decoder models, reinforcing the suitability of sequence-to-sequence architectures.

4.4. Dialect-Aware Data Augmentation

To analyze dialect coverage and explore robustness, We used the MARBERTv2 Arabic Written Dialect Classifier to label dialects in the training data. The original distribution over 6,263 instances was skewed toward MSA (Modern Standard Arabic): MSA (49.77%), GLF (19.99%), EGY (13.09%), LEV (10.49%), and MAGHREB (6.66%).

To improve balance, we performed targeted augmentation on underrepresented dialects (EGY, LEV, MAGHREB). The dataset was split into minority and majority subsets, and lightweight label-preserving transformations were applied to the minority samples, including tatweel removal, controlled elongation, punctuation variation, and optional emoji insertion. One augmented instance was generated per selected sample while keeping the target text unchanged.

The augmented dataset grew to 8,157 instances with a more balanced distribution: MSA (38.22%), EGY (20.11%), LEV (16.11%), GLF (15.35%), and MAGHREB (10.21%). However, this strategy did not yield consistent gains.

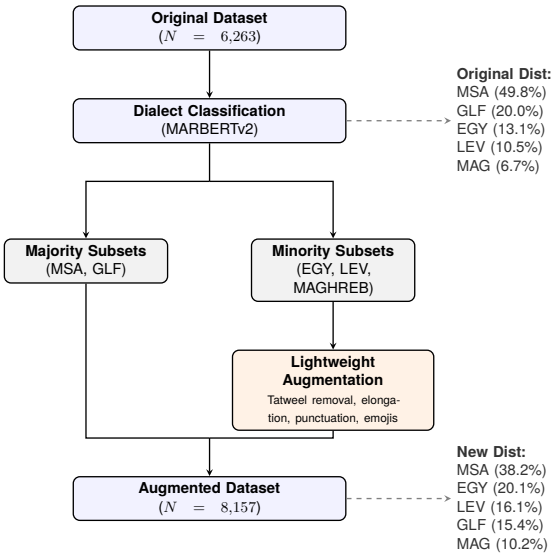

Figure 1: Overview of the dialect-aware data augmentation pipeline. The original dataset is classified and split, applying targeted lightweight transformations only to underrepresented dialects to achieve a more balanced distribution.

4.5. Training Strategy

Encoder–decoder models are fine-tuned using standard sequence-to-sequence cross-entropy loss, enabling explicit learning of sentiment-inversion mappings. Fine-tuned Arabic models provide the strongest baseline performance.

4.6. Ensemble Inference

Our final submission ensembles AraT5v2 and AraBART. Both models are trained independently, and during inference each input sentence is processed by both models to produce candidate rewrites.

CAMeLBERT-DA is then used as a sentiment judge, selecting the candidate whose predicted polarity best matches the target. This selection-based ensemble improves sentiment correctness while maintaining fluency.

5. Results and Discussion

We evaluate generation quality using BLEU,chrF, Loss and Accuracy. BLEU measures n-gram precision between generated and reference sentences, reflecting lexical fidelity. Together, these metrics provide a complementary assessment of semantic consistency and surface-level accuracy in the sentiment swap task.

To explicitly verify polarity inversion, we additionally employ CAMeLBERT-DA, a dialect-aware BERT-based Arabic sentiment classifier. The model encodes sentences using bidirectional self-attention and predicts sentiment from contextualized representations, providing a sentiment-oriented evaluation signal that complements reference-based metrics.

Table 1 shows that fine-tuned Arabic encoder–decoder models substantially outperform multilingual and decoder-only approaches for Arabic sentiment swap. Overall, models pretrained specifically for Arabic demonstrate stronger sentiment inversion and better semantic preservation, confirming the importance of language-specific pretraining for controlled generation in morphologically rich languages such as Arabic.

Model	Loss	BLEU	chrF	Acc.
AraT5v2	**0.99**	**42.71**	**65.0**	**74.61**
AraBART	0.82	42.82	65.53	69.66
mT5	0.998	54.44	40.02	53.5
GPT-J-6B	0.32	53	33.7	44

Table 1: Performance comparison for each model individually on the MA'aks test set.

5.1. Encoder–Decoder vs. Decoder-Only Models

Our experiments show a clear advantage of encoder–decoder architectures over decoder-only generation. While GPT-J-6B produces fluent outputs, it struggles to reliably invert sentiment and preserve meaning, resulting in weaker overall performance. In contrast, sequence-to-sequence models benefit from explicit input–output alignment, making them better suited for controlled sentiment rewriting.

5.1.1. Multilingual vs. Arabic Encoder–Decoder Models

Within encoder–decoder approaches, Arabic-focused models consistently outperform the multilingual mT5. Despite its broad pretraining, mT5 fails to capture fine-grained sentiment transformations in Arabic and exhibits substantially weaker chrF and accuracy scores. These results indicate that multilingual pretraining alone is insufficient for sentiment-controlled generation in morphologically rich languages.

In contrast, AraT5v2 and AraBART demonstrate strong and stable performance across metrics. Their Arabic-specific pretraining enables more precise manipulation of sentiment-bearing expressions. Given these findings, our subsequent experiments and system design focus primarily on Arabic-pretrained models rather than multilingual alternatives.

5.2. Ensemble Performance

We further evaluate the proposed ensemble strategy combining AraT5v2 and AraBART. When used individually under fine-tuned settings, AraT5v2 achieves BLEU 42.71, chrF 65.0, and sentiment preservation 0.7461, while AraBART obtains BLEU 42.82, chrF 65.53, and sentiment preservation 0.6966. Although AraBART slightly outperforms AraT5v2 in lexical metrics, AraT5v2 provides stronger sentiment preservation.

By leveraging CAMeLBERT-DA as a selection mechanism during inference, the ensemble effectively combines the complementary strengths of both models. This results in our best-performing system, achieving BLEU 43.0, chrF 65.36, and sentiment preservation 0.7554, which secured first place in AraSentEval 2026 Subtask 2. These findings confirm that selection-based ensembling improves sentiment correctness while maintaining high generation quality.

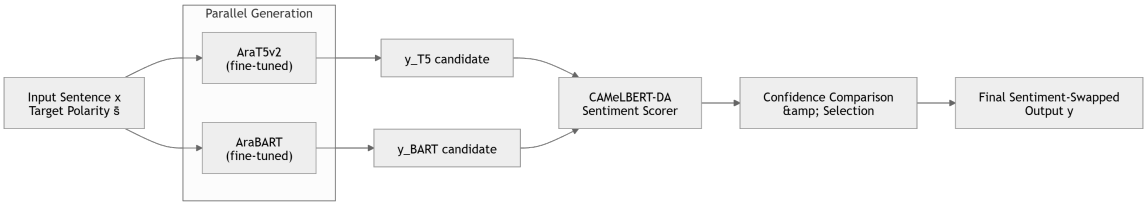

Figure 2: Ensemble Parallel AraT5v2/AraBART generation with CAMeLBERT-DA–based selection.

6. Error Analysis

6.1. Major Failure Modes

Despite strong overall performance, several systematic failure patterns remain. The most prominent source of error is dialect imbalance. Although we experimented with targeted augmentation, the approach did not fully resolve skew in dialect coverage. Performance degradation is most visible on underrepresented varieties and on sentences exhibiting dialectal ambiguity, particularly between MSA and Gulf Arabic, where lexical overlap complicates reliable polarity manipulation.

The model also struggles with informal phenomena such as emojis and idiomatic expressions, where sentiment is conveyed implicitly rather than through explicit polarity markers. In such cases, the system often preserves fluency but fails to correctly invert sentiment.

6.2. Qualitative Examples

Representative challenging cases include:

- **Arabic (original):** مر رهيب و رائع وطعم اللذة يستحق مليون نجمة
  Transliteration: mar rahīb wa rāʾiᶜ wa ṭaᶜm al-ladhdha yastaḥiq milyūn najma.
  Translation: ‘It is amazing and wonderful, and the delicious taste deserves a million stars.’

- **Arabic (original):** تطبيق رائع والطلب منه كتير سهل
  Transliteration: taṭbīq rāʾiᶜ wa al-ṭalab minhu katīr sahl.
  Translation: ‘A wonderful app, and ordering from it is very easy.’

These examples illustrate failures involving implicit sentiment cues, emoji-driven polarity, and dialect-sensitive phrasing.

6.3. Root Cause Analysis

The primary limitation stems from dialect distribution mismatch and cross-dialect similarity. When dialect boundaries are blurred (e.g., MSA vs. Gulf), the system may misidentify sentiment-bearing spans or apply inappropriate rewrites. Our augmentation strategy improved dialect balance statistically but did not sufficiently enhance model robustness to these ambiguities.

6.4. Future Work

Future work can extend this study in several directions. First, more advanced fine-tuning strategies such as multi-stage fine-tuning and curriculum learning could gradually expose models to increasingly complex sentiment transformations. Second, stronger augmentation methods such as dialect-controlled paraphrasing and back-translation may better address dialect imbalance. Third, synthetic data generation using large language models could help create larger distribution-matched corpora for robust fine-tuning. Finally, alternative objectives beyond standard cross-entropy, including sentiment-aware or contrastive losses, may better enforce polarity correctness while preserving semantic content.

7. Conclusion

In this paper, we study Arabic sentiment swap as a controlled generation task requiring precise polarity inversion while preserving meaning and fluency. Experiments on the MA’AKS dataset show that reliable sentiment manipulation in Arabic benefits more from language-specific supervised learning than from purely multilingual approaches.

Our results demonstrate that Arabic encoder–decoder models, particularly AraT5v2 and AraBART, outperform multilingual and decoder-only baselines. We further propose a selection-based ensemble that leverages their complementary strengths using CAMeLBERT-DA as a sentiment-aware judge. The proposed system achieves state-of-the-art performance and ranks first in AraSentEval 2026 Subtask 2.

Error analysis reveals persistent challenges, including dialect imbalance, implicit sentiment cues, and emoji-driven expressions. Overall, this work advances sentiment-controlled generation in Arabic and establishes strong baselines for future research on style transfer in morphologically rich and dialectally diverse languages.

8. Bibliographical References

M. Elmadany, A. Elbehery, and M. Abdelrahman. 2021. Arabart: Transformer-based model for arabic text generation. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Saad Ezzini, Shadi Abudalfa, Maram Alharbi, Salmane Chafik, Hamzah Luqman, Mo El-Haj, Paul Rayson, Samaher Alghamdi, and Reem Alotaibi. 2026. AraSentEval: A shared task on sentiment analysis and swapping in arabic. In *Proceedings of the 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

A. Inoue, M. Alsharif, and H. Mubarak. 2021. Camelbert: A collection of pretrained transformer models for arabic. In *Proceedings of the ACL Workshop on NLP for Arabic*.

A. Nagoudi, M. El-Beltagy, and A. Abdelali. 2022. Arat5: Text-to-text transformers for arabic language generation. In *Proceedings of the ACL Workshop on NLP for Arabic*.

B. Wang and A. Komatsuzaki. 2021. Gpt-j-6b: A 6 billion parameter autoregressive language model. Technical report, EleutherAI.

L. Xue, N. Constant, A. Roberts, et al. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of NAACL*.

9. Language Resource References

Raed Mughaus, Shadi Abudalfa, Hamzah Luqman, Fahad Abdu, Mohammed AlAli, Nawaf Al-Dowayan, and Ahmed Abdelali. 2026. Ma'aks: manually-curated parallel dataset for arabic text sentiment swap. *Language Resources and Evaluation*, 60(1):1.