L3IA at AraSentEval Shared Task: LLM-Based Multi-Step Pipeline for Arabic Sentiment Swap

Hamza Alami, Mohamed M'haouach, Kaouthar Elyoussoufi, Abdessamad Benlahbib
L3IA Laboratory, Faculty of Sciences Dhar El Mahraz,
Sidi Mohamed Ben Abdellah University, Fez, 30003, Morocco
{hamza.alami5, mohamed.mhaouach, kaouthar.elyoussoufi,
abdessamad.benlahbib}@usmba.ac.ma

Abstract

This paper describes our system submitted to the AraSentEval 2026 Shared Task, Subtask 2: Arabic Sentiment Swap. The task requires rewriting Arabic sentences to invert their sentiment polarity while preserving the core meaning. We propose a multi-step pipeline approach that uses large language models (LLMs). Our method decomposes the sentiment inversion problem into three stages: (1) sentiment expression extraction, where the model identifies all sentiment-bearing words and phrases in the input sentence; (2) opposite expression generation, where each identified expression is replaced by its semantic opposite; and (3) sentence reconstruction, where the final output is assembled to ensure grammatical correctness and natural fluency. Our system achieves 74.3% sentiment style accuracy, 27.22 BLEU, and 55.04 chrF on the official test set.

Keywords: Sentiment Analysis, Arabic sentiment swap, NLP

1. Introduction

Sentiment analysis is a well-studied area in natural language processing (NLP). However, the reverse task — transforming a sentence's sentiment from positive to negative or vice versa — remains a challenging and less explored problem. This task, known as sentiment style transfer or sentiment swap, requires models to go beyond classification and actually generate text with the opposite polarity while keeping the original meaning intact (Jin et al., 2022; Hu et al., 2022; Toshevska and Gievska, 2022).

Arabic presents additional challenges for this task. It is a morphologically rich language with a complex system of roots, patterns, and affixes. Moreover, the Arab world exhibits significant dialectal variation, with Egyptian, Gulf, Levantine, and Maghrebi dialects differing from Modern Standard Arabic (MSA) in vocabulary, grammar, and usage (Younes et al., 2020; Joshi et al., 2025). Social media content, which forms the basis of the dataset used in this shared task, is especially challenging because it mixes dialects, uses informal spelling, and includes emojis and code-switching (Hajbi et al., 2022).

The AraSentEval 2026 Shared Task (Ezzini et al., 2026) addresses these challenges through two subtasks. Subtask 1 focuses on sentiment detection in Arabic dialects, while Subtask 2 — the focus of this paper — targets sentiment swap. Given an Arabic sentence and its source polarity, the goal is to produce a new sentence that preserves the core meaning but expresses the opposite sentiment. The dataset is built from MA'AKS (Mughaus et al., 2026), a manually curated parallel dataset for Arabic text sentiment swap, containing sentence pairs in MSA and multiple Arabic dialects (Egyptian, Gulf, Levantine, and Maghrebi).

Recent advances in large language models (LLMs) have shown strong capabilities in text generation, including style transfer tasks (Toshevska and Gievska, 2025; Zhang and Tang, 2025). Rather than fine-tuning a sequence-to-sequence model on the training data, we propose a prompt-based approach that leverages the linguistic knowledge already encoded in LLMs. Our method breaks the sentiment swap task into three interpretable sub-tasks — analysis, replacement generation, and reconstruction — each handled by a dedicated prompt. This decomposition allows the model to focus on one aspect at a time, leading to more controlled and accurate sentiment inversion.

Our main contributions are:

- A **multi-step LLM pipeline** for Arabic sentiment swap that decomposes the task into sentiment expression extraction, opposite generation, and sentence reconstruction.
- An **asynchronous batch processing** system that allows parallel API calls for efficient large-scale inference.
- An analysis of the strengths and limitations of prompt-based LLM approaches for Arabic text style transfer.

2. Methodology

Given an input Arabic sentence s with sentiment polarity p in {positive, negative}, the task is to gen-

Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks, pages 274–277
May 11, 2026.

erate an output sentence $\hat{s}$ such that: (1) $\hat{s}$ has the opposite polarity, and (2) $\hat{s}$ preserves the core meaning and structure of s as much as possible. The system is evaluated on sentiment style accuracy (whether $\hat{s}$ has the correct target polarity), and content preservation (BLEU and chrF between s and $\hat{s}$).

Our system uses a three-stage pipeline, where each stage is handled by a separate LLM call with a dedicated prompt. Figure 1 depicts the overall flow of our Arabic sentiment swapper.

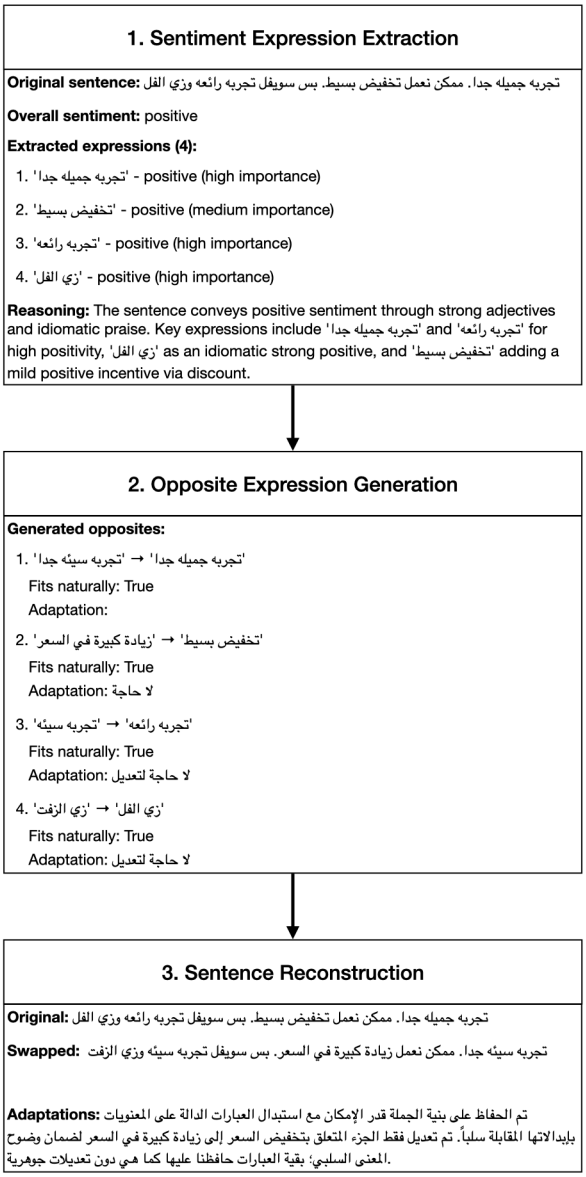

Figure 1: The overall flow of our Arabic sentiment swapper

We use OpenAI's chat models accessed through the LangChain framework. Each step produces structured JSON output that is validated using Pydantic data models to ensure consistency and reliability.

2.1. Sentiment Expression Extraction

The first step identifies all sentiment-bearing expressions in the input sentence. We prompt the LLM to act as an Arabic language and sentiment analysis expert. The model is asked to detect:

- Adjectives and descriptive phrases (e.g., "جميلة" meaning beautiful)
- Verbs with emotional connotations (e.g., "أحب" meaning I love)
- Adverbs that modify sentiment intensity (e.g., "جداً" meaning very)
- Idiomatic expressions (e.g., "زي الفل" meaning great/perfect)
- Negations that affect sentiment (e.g., "مش" meaning not)

For each identified expression, the model returns its text, character position in the sentence, sentiment influence (positive or negative), and importance level (high, medium, or low). The output is parsed into a structured *SentimentAnalysis* object containing the original sentence, overall sentiment, list of expressions, and the model's reasoning.

2.2. Opposite Expression Generation

For each sentiment expression identified in Step 1, the model generates an opposite expression. The prompt includes the full original sentence as context to ensure that the replacement fits semantically and grammatically. The model is instructed to:

- Generate an expression that conveys the opposite sentiment
- Ensure grammatical agreement (gender, number, case) with the surrounding context
- Indicate whether the replacement fits naturally in the same position
- Suggest any necessary adaptations if it does not fit directly

Each replacement is returned as an *OppositeExpression* object containing the original and opposite expressions, a naturalness flag, and adaptation notes.

2.3. Sentence Reconstruction

The final step takes the original sentence and all planned replacements and produces the output sentence. The reconstruction prompt instructs the model to:

- Apply all expression replacements from Step 2.2

- Maintain the original sentence structure and word order as much as possible

- Ensure grammatical correctness in the final output

- Make only minimal adaptations when necessary for naturalness

The model returns a *SwappedSentence* object containing the original sentence, the swapped sentence, the list of replacements made, and a description of any adaptations.

2.4. Model Configuration and Batch Processing

We use OpenAI's GPT 5 Nano model in our work. To process the test set efficiently, we implement an asynchronous batch processing system using Python's *asyncio*. This system processes multiple samples in parallel.

3. Experimental Results

3.1. Dataset

The dataset is provided by the AraSentEval 2026 shared task organizers and is based on the MA'AKS corpus (Mughaus et al., 2026). The training and validation sets contain aligned sentence pairs with source polarity labels. The test set contains only the input sentences without polarity labels or reference outputs. The data includes sentences in MSA and multiple Arabic dialects.

3.2. Evaluation Metrics

The official evaluation uses three metrics, namely, Sentiment Style Accuracy is the percentage of outputs classified with the correct target polarity by the *CAMeL-Lab/bert-base-arabic-camelbert-da-sentiment* model (Inoue et al., 2021), BLEU (Papineni et al., 2002), and chrF (Popović, 2015) measure how much of the original sentence's content is retained in the output. Higher scores indicate better meaning preservation but may also indicate insufficient sentiment change.

3.3. Results and Discussion

Table 1 shows our system's performance on the official test set. To better understand our system's behavior, we analyze several output examples. Table 2 shows representative cases.

Table 1: Obtained results

Model	Sentiment Accuracy (%)	BLEU	chrF
OpenAI GPT-5-Nano	74.3%	27.22	55.04

We observe several patterns in our system's outputs. **1- Expression-level precision**: The system correctly identifies and replaces sentiment-bearing expressions while leaving neutral content unchanged, which helps preserve the original meaning. **2- Dialect awareness**: The model handles dialectal expressions well (e.g., "زي الفل" becomes "زي الزفت", "مفيش" becomes "فيه"), showing that the LLM has good knowledge of Arabic dialects. **3- Structural preservation**: By constraining the reconstruction step to follow the original sentence structure, the system produces outputs that are close to the input, which benefits BLEU and chrF scores.

We observe also various limitations. **1- Error propagation**: Errors in step 2.1 (missing a sentiment expression) propagate to later steps, resulting in incomplete sentiment inversion. **2- Over-generation**: In some cases, the reconstruction step introduces words or phrases not present in the original sentence, reducing content preservation scores. **3- Emoji handling**: The system sometimes struggles with emojis, replacing them with text descriptions rather than opposite emojis. **4- Multiple API calls**: Each sample requires three separate API calls, making the system slower and more expensive than a single-model approach.

When comparing with fine-tuning approaches, our prompt-based approach offers several advantages over fine-tuning: it requires no training data or GPU resources, can be deployed immediately, and benefits from the LLM's pre-existing knowledge of Arabic. However, fine-tuned models may achieve better content preservation since they can learn the specific transformation patterns present in the training data.

4. Conclusion

We presented a multi-step LLM pipeline for Arabic sentiment swap as part of the AraSentEval 2026 Shared Task 2. Our approach decomposes the problem into three interpretable stages — sentiment expression extraction, opposite generation, and sentence reconstruction — each handled by a dedicated prompt. The system uses structured output parsing for reliability and asynchronous processing for scalability.

Our results show that prompt-based LLM approaches can perform sentiment inversion in Arabic, including dialectal text. The multi-step design provides transparency and interpretability, as each

Table 2: Example Outputs

Input	Output	Notes
كل ما انزل الابلكيشن وادخل رقم التليفون لم يتم الطلب ... ومفيش تواصل مع اي. مسئول	كل ما انزل الابلكيشن وادخل رقم التليفون تم الطلب بنجاح، وقد استجاب للطلب وفيه تواصل مع أي حد مسئول	Negation removal + positive restructuring
انا نزلت الابلكيشن ودخلت رقم الموبايل وبعد كدة جيت ادخل المحافظة قالي حصل خطأ !	انا نزلت الابلكيشن ودخلت رقم الموبايل وبعد كدة جيت ادخل المحافظة قالي تم بنجاح !	Minimal change, good preservation

intermediate step can be inspected and debugged independently.

For future work, we plan to: 1- explore combining LLM-based approaches with fine-tuned models in a hybrid system, 2- investigate few-shot prompting with examples from the training set to improve quality, and 3- experiment with newer and more capable models.

References

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.

Soufiane Hajbi, Younes Chihab, Rachid Ed-Dali, and Redouan Korchiyne. 2022. Natural language processing based approach to overcome arabizi and code switching in social media moroccan dialect. In *Advances in Information, Communication and Cybersecurity*, pages 57–66, Cham. Springer International Publishing.

Zhiqiang Hu, Roy Ka-Wei Lee, Charu C. Aggarwal, and Aston Zhang. 2022. Text style transfer: A review and experimental evaluation. *SIGKDD Explor. Newsl.*, 24(1):14–45.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2025. Natural language processing for dialects of a language: A survey. *ACM Comput. Surv.*, 57(6).

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Martina Toshevska and Sonja Gievska. 2022. A review of text style transfer using deep learning. *IEEE Transactions on Artificial Intelligence*, 3(5):669–684.

Martina Toshevska and Sonja Gievska. 2025. Llm-based text style transfer: Have we taken a step forward? *IEEE Access*, 13:44707–44721.

Jihene Younes, Emna Souissi, Hadhemi Achour, and Ahmed Ferchichi. 2020. Language resources for maghrebi arabic dialects'nlp: a survey. *Language Resources and Evaluation*, 54(4):1079–1142.

Tianshan Zhang and Hao Tang. 2025. Style transfer: A decade survey.