AraSentEval 2026: A Shared Task on Sentiment Analysis and Swapping in Arabic

Saad Ezzini[1], Shadi Abudalfa[1], Maram Alharbi[2], Salmane Chafik[3],
Hamzah Luqman[1], Mo El-Haj[2,4], Paul Rayson[2,4], Samaher Alghamdi[2], Reem Alotaibi[5]
[1]King Fahd University of Petroleum & Minerals, KSA
[2]Lancaster University, UK
[3]Mohammed VI Polytechnic University, Morocco
[4]VinUniversity, Vietnam
[5]King Abdulaziz University, KSA

Abstract

Sentiment analysis is a fundamental problem in Natural Language Processing (NLP). Standard sentiment classification for the Arabic language remains challenging due to the high volume of dialectal Arabic. To advance research in this area, this paper proposes the Shared Task on Sentiment Analysis and Swapping in Arabic (AraSentEval), organized as part of the OSACT7 Workshop at LREC 2026. This shared task consists of two subtasks: Subtask 1 focuses on multi-class and multi-dialect sentiment analysis, requiring models to identify sentiment polarity across various Arabic dialects. Subtask 2 introduces a generative task for Arabic sentiment swap, challenging models to invert sentiment polarity while preserving core semantics. In this overview paper, we present the motivation, dataset creation, and summarize the main findings from participating models.

Keywords: Arabic NLP, Sentiment Analysis, Text Style Transfer, Shared Task

1. Introduction

Sentiment analysis remains a key task in Natural Language Processing (NLP), with critical real-world applications ranging from social media monitoring to customer feedback analysis. While significant progress has been made for many languages, and multiple types have been defined (document, sentence, and feature/aspect level), the Arabic language presents unique challenges due to its rich morphology and extensive dialectal variation. Most user-generated content, the primary source for sentiment data, is written in dialectal Arabic, which is often under-resourced and differs significantly from Modern Standard Arabic (MSA).

To address these challenges and foster innovation in Arabic NLP, we propose a novel Shared Task on Sentiment Analysis and Swapping in Arabic (AraSentEval), hosted at the OSACT7 Workshop at LREC 2026[1]. This task is designed to move beyond standard sentiment classification by evaluating both the understanding and generation of sentiment in diverse Arabic contexts. AraSentEval comprises two distinct but complementary subtasks:

- **Subtask 1: Arabic Dialect Sentiment Analysis**: A multi-class classification task focused on identifying the sentiment (positive, negative, or neutral) of texts written in various Arabic dialects. This subtask emphasizes the need for robust models that can handle the lexical and syntactic diversity of the Arab world.
- **Subtask 2: Arabic Sentiment Swap**: A generative task where participants' systems must rewrite a given sentence to invert its sentiment polarity (e.g., positive to negative) while preserving its core meaning. This challenges models to go beyond surface-level cues and demonstrate a deeper grasp of semantics and syntax.

2. Motivation

The motivation for AraSentEval is grounded in both linguistic and methodological gaps in current Arabic NLP research.

First, despite sustained interest in sentiment analysis for Arabic, there remains a clear shortage of robust benchmarks that adequately capture dialectal diversity. Earlier shared tasks and evaluation campaigns (El-Beltagy et al., 2017; Rosenthal et al., 2017) have advanced sentiment classification for Arabic, yet most resources are either limited to Modern Standard Arabic (MSA) or focus on a narrow set of dialects and domains (El-Haj, 2026, 2020). In practice, however, the majority of user-generated content, including reviews, social media posts, and online discussions, is written in dialectal Arabic. These dialects exhibit substantial lexical, morphological, and syntactic variation, which often leads to performance degradation when models trained on MSA are applied to real-world data. Prior work has highlighted the fragmentation of Arabic sentiment resources and the difficulty of transferring models across dialects (Nwesri et al., 2025; Alwakid et al.,

[1] https://ezzini.github.io/AraSentEval/

Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks, pages 256–261

2022), reinforcing the need for more representative evaluation benchmarks. By introducing a balanced, multi-dialect benchmark informed by recent benchmarking efforts (Alharbi et al., 2025c), AraSentEval seeks to promote the development of sentiment models that generalise across dialects and better reflect authentic language use in the Arab world.

Second, the task aims to extend evaluation beyond classification into controlled text generation. While sentiment classification has reached a relatively mature stage in Arabic NLP, generative sentiment manipulation remains underexplored. Sentiment swapping, framed as a form of style transfer, requires systems to invert polarity while preserving core propositional content. This problem has been investigated in English (Shen et al., 2017; Li et al., 2018), where style transfer research has demonstrated the complexity of disentangling content from sentiment and other stylistic attributes. Comparable resources and systematic evaluation frameworks are scarce for Arabic, limiting progress in controllable generation (Abudalfa et al., 2024; Abdu et al., 2025). The limited availability of high-quality parallel datasets has constrained progress in this area. By introducing a dedicated sentiment swap subtask, AraSentEval encourages research on semantic preservation, polarity control, and evaluation methodologies for Arabic natural language generation. Such capabilities have practical implications for data augmentation, controllable dialogue systems, and adaptive content generation, particularly in low-resource and dialectally diverse settings.

By combining a dialect-aware classification task with a challenging generative task, AraSentEval offers a unified benchmark for both sentiment understanding and sentiment manipulation in Arabic. This dual focus is intended to stimulate research that moves beyond surface-level polarity detection towards more semantically grounded and controllable language technologies for Arabic.

3. Data Collection and Creation

The datasets for both subtasks were collected and annotated prior to the task launch. Figure 1 illustrates the dataset creation process for both subtasks.

3.1. Subtask 1: Dialect Sentiment Analysis

The dataset for this task, Multi-Dialect-Sent (MDS-3), currently consists of 3,000 sentences annotated for three-class sentiment (positive, negative, neutral). The dataset is balanced across four major dialects: Moroccan, Egyptian, Jordanian, and Saudi. The data was sourced from hotel reviews from the

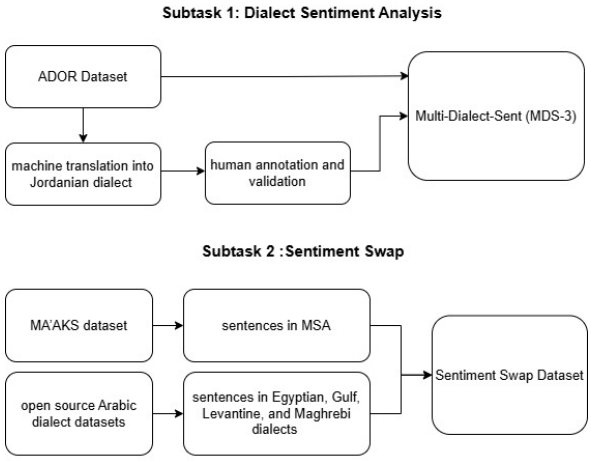

Figure 1: Data creation process for both subtasks.

ADOR dataset (Alharbi et al., 2025b), then translated and annotated by native speakers of each dialect. The sentences are parallel across dialects, with balanced sentiment representation across all included dialects. An initial version of the dataset was verified in a recently organised shared task Ahasis (Alharbi et al., 2025a) at RANLP-2025.

The dataset was split into 80% for training and 20% for testing. The split was performed based on sentence IDs to ensure that parallel sentences remain in the same set, to prevent potential data leakage between the training and test sets.

3.2. Subtask 2: Sentiment Swap

The data employed in this subtask is primarily derived from the MA'AKS dataset (Mughaus et al., 2026), a manually constructed parallel resource designed for Arabic sentiment swap. This dataset contains 5,000 paired sentences in MSA. Within each pair, both sentences address the same underlying subject matter but express contrasting sentiment orientations—one conveying a positive stance and the other a negative one. The dataset was assembled from publicly accessible resources and underwent careful manual annotation and cross-checking to guarantee linguistic fluency, semantic consistency, and precise sentiment transformation.

For experimental purposes, MA'AKS dataset is partitioned into three subsets: 4,000 pairs for training, 500 for validation, and 500 for testing. It has been evaluated using multiple advanced Large Language Models (LLMs), such as AceGPT, JAIS, and Llama-3, under different configurations including zero-shot prompting, few-shot prompting, and supervised fine-tuning. These evaluations demonstrate both the robustness of the dataset and the practicality of the sentiment-swapping task.

Beyond MA'AKS dataset, the subtask also incorporates additional data representing several

Arabic dialects, namely Egyptian, Gulf, Levantine, and Moroccan varieties. These dialectal examples were sourced from other open datasets and subsequently annotated and verified with the assistance of open-source LLMs. Following this integration and validation process, the final dataset for the subtask comprises 6,263 MSA and dialectal pairs in the training split, 1,315 pairs in the development split, and 646 dialect-only pairs in the test split. The test set was intentionally restricted to dialectal Arabic to rigorously evaluate the systems' generalization abilities to dialectal variations, excluding the relatively easier MSA content.

4. Task Description and Evaluation

In this section, we describe in detail the subtasks that define our shared task. Participants were allowed to participate in either one or both subtasks. The official evaluation was conducted on the **CodaBench** platform, an open-source platform designed to organize Artificial Intelligence benchmarks by defining tasks, including datasets, submission pipelines, and evaluation metrics. It provides an automated and transparent framework for evaluating system submissions.

4.1. Subtask 1: Arabic Dialect Sentiment Analysis

In this subtask, participants are required to build a system that classifies the sentiment expressed in sentences written in four different and linguistically challenging Arabic dialects: Moroccan (Darija), Egyptian, Jordanian, and Saudi. The goal of this task is to evaluate the ability of models to handle dialectal variation in Arabic while accurately identifying the sentiment conveyed in each sentence.

Formally, the objective is to design a sentence-level classifier C such that, given an input sentence x, the system predicts a sentiment label S, where:

$$S = C(x), \quad S \in \{\text{Positive}, \text{Neutral}, \text{Negative}\}.$$

- **Input**: A sentence written in Arabic using one of the target dialects (Moroccan, Egyptian, Jordanian, or Saudi).
- **Output**: A sentiment label selected from the set {Positive, Neutral, Negative}.
- **Evaluation Metric**: Systems are evaluated using the **Macro F1-score**. This metric is chosen because it gives equal importance to all classes and is therefore more suitable in scenarios where class distributions may be imbalanced.

4.2. Subtask 2: Arabic Sentiment Swap

In the second subtask, participants are expected to design models able to generate a new sentence that expresses the opposite sentiment of a given sentence while preserving its original semantic content. The input sentences are written in either Modern Standard Arabic (MSA) or in one of several Arabic dialects, including Egyptian, Gulf, Levantine, and Moroccan. This task evaluates a model's ability to perform controlled sentiment transformation while maintaining semantic fidelity.

Formally, the objective is to train a model M capable of performing sentiment-controlled text generation. Given an input sentence x expressing a sentiment S, the model should generate a new sentence x' that conveys the opposite sentiment S', while preserving the core meaning and topic of the original sentence, such that:

$$x' = M(x), \quad S' = \text{Opposite}(S)$$

where $S, S' \in \{\text{Positive}, \text{Negative}\}$ and $S' \neq S$.

- **Input**: An Arabic sentence along with its original sentiment polarity (e.g., "هذا المطعم رائع للغاية", Positive).
- **Output**: A rewritten Arabic sentence that preserves the main meaning of the input while expressing the opposite sentiment polarity (e.g., "هذا المطعم سيء للغاية").
- **Evaluation Metrics**: The generated outputs are evaluated using both sentiment accuracy and text similarity metrics to ensure correct sentiment transfer while maintaining semantic consistency.
  - **Sentiment Style Accuracy**: A pre-trained sentiment classifier is used to measure the percentage of generated sentences that correctly express the target polarity.
  - **BLEU and chrF**: These metrics are used to evaluate the degree of lexical and character-level similarity between the generated sentence and the reference sentence, providing an indication of semantic preservation.

5. Participant Systems

A total of 15 teams participated in the shared task, of which 9 submitted system description papers for AraSentEval. Below is a brief summary of the methodologies proposed by the participating teams, mapping their team names to their corresponding leaderboard usernames, their chosen subtask, and their reported scores and rankings:

- **musj1984** (Username: *musj1984*): Participated in **Subtask 1**. They adapted a Mixture of Experts (MoE) framework to handle Arabic three-class sentiment classification by combining two QLoRA fine-tuned models (AraBERT v2 and BERT-Large Arabic) using a learned gating network. Training data was enriched via LLM-based augmentation. As reported in their paper, they achieved a Macro F1-score of 0.9423, ranking **1st** overall in the subtask.

- **TTLab** (Username: *alighabusaleh*): Participated in **Subtask 1**. They proposed SARF, a multi-view architectural framework integrating surface-level context with stemmed and rooted morphological perspectives via a shared MARBERTv2 encoder. Utilizing a hybrid BERT-CNN-BiLSTM-Attention architecture, they reported a competitive Macro F1-score of 0.9263, ranking **2nd** out of 15 participating teams.

- **CasbAI** (Username: *abdelazizchaima*): Participated in **Subtask 1**. They designed a Transformer-based approach leveraging MARBERT with a multi-seed ensemble strategy and Stochastic Weight Averaging (SWA). Through dialectal synonym replacement and Test-Time Augmentation (TTA), they reported achieving a Macro F1-score of 0.8462, ranking **3rd** among participating teams.

- **LinguArabic** (Username: *elham42*): Participated in **Subtask 1**. They fine-tuned MARBERT using a multi-stage preprocessing pipeline that included text normalization, dialect-aware lexical mapping, and confidence-based prediction adjustment. They reported achieving a Macro F1-score of 0.8333 on the official evaluation set (placing them **5th** closely on the leaderboard).

- **BDSI** (Username: *stdai*): Participated in **Subtask 1**. They developed a multi-model ensemble combining AraBERTv2 and CAMeLBERT with supervised contrastive learning. With a pipeline integrating dialect-aware preprocessing and rule-based post-processing, they reported a Macro F1-score of 0.83 (matching the **6th** placed system score of 0.8301).

- **WIT'INNOV & L3IA** (Username: *beatchalvador*): Participated in **Subtask 1**. They conducted a benchmarking study of five Transformer-based architectures, emphasizing a text normalization procedure to unify orthographic variations. They reported achieving an internal validation F1-score of 94.92%, while their official submission scored a Macro F1-score of 0.7596, placing them **13th** in the leaderboard.

- **Unnamed Team** (Username: *yumnahamdy*): Participated in **Subtask 2**. They investigated encoder-decoder and multilingual approaches for Arabic sentiment swap, leveraging an ensemble of AraBART and AraT5v2 with dialect-aware data augmentation. In their paper, they reported ranking **1st** in Subtask 2, achieving a BLEU score of 43.0, chrF of 65.36, and a sentiment preservation accuracy of 0.7554.

- **L3IA** (Username: *beatchalvador*): Participated in **Subtask 2**. They Addressed the task using a multi-step pipeline employing large language models (LLMs) to extract sentiment expressions, generate opposites, and reconstruct the sentence. They reported achieving 74.3% sentiment style accuracy, 27.22 BLEU, and 55.04 chrF, cleanly matching the **2nd** place leaderboard position.

- **VGU-M.Tech-AI** (Username: *asbichi362*): Participated in **Subtask 2**. They fine-tuned a multilingual T5 (mT5) model (ASBN-MT5) in a sequence-to-sequence manner. They reported conversion rates of 59.5% for positive to negative and 58.5% for negative to positive polarity swaps while achieving a BLEU score of 30.79 and a ChrF score of 54.08 on the official evaluation.

6. Results

Dialect Sentiment Analysis. Table 1 presents the results of the submissions for the Arabic Dialect Sentiment Analysis subtask, evaluated using the Macro F1-score. As shown in the table, the Musj1984 system achieved the highest F1-score of 0.9423, followed by TTLab with an F1-score of 0.9263. CasbAI ranked third, achieving an F1-score of 0.8462. Notably, the top-performing teams employed BERT-based variants as backbone encoder models. These models outperformed systems such as L3IA, which relied on LLMs for this task. This performance difference can be largely attributed to the Arabic training data used to train Arabic BERT variants, as well as the strong representation capabilities of BERT models, which make them well-suited for classification tasks. Additionally, systems that used ensemble approaches, such as Musj1984 and TTLab, significantly outperformed single-model approaches. This improvement can be attributed to the ability of ensemble methods to enhance prediction reliability by combining multiple model outputs, thereby increasing overall classification confidence.

Sentiment Swap. Table 2 presents the results for the Arabic Sentiment Swap subtask based on the Sentiment Style Accuracy, along with the generation quality metrics, BLEU, and chrF. As shown

#	Team/Participant Name	Macro F1-Score
1	Musj1984	0.9423
2	TTLab	0.9263
3	CasbAI	0.8462
4	University of Tripoli	0.8429
5	LinguArabic	0.8333
6	BDSI	0.8301
7	Mohammadnabulsi	0.8205
8	Gehadkamel	0.8205
9	Amani_Sh	0.8045
10	Tareqkhaled	0.7981
11	Mary-Sussex-ai	0.7981
12	Sarahmeid	0.7885
13	L3IA	0.7430
14	Amjad_Awad	0.6506
15	Astral_Fate	0.3686

Table 1: Results of the submitted systems for Subtask 1: Arabic Dialect Sentiment Analysis.

in the table, the leaderboard results for this task show clear differences in how well systems balance sentiment transformation and content preservation. The Yumnahamdy team ranks first with the highest sentiment style accuracy (0.7554), along with the best BLEU (43.00) and ChrF (65.36) scores. These results indicate a strong ability to change the sentiment while preserving the original meaning and structure. The L3IA team follows closely in second place with a similar sentiment accuracy (0.7430) but noticeably lower BLEU. These results indicate that the proposed model prioritizes sentiment transfer slightly more than exact wording. In contrast, the VGU-M.Tech-AI team performs poorly in sentiment accuracy (0.2368) despite having relatively better BLEU and ChrF than some competitors. This difference between the two metrics indicates that the model largely preserves the original text but does not effectively modify the sentiment. Overall, the results highlight that the key challenge of this task lies in achieving a balance between accurate sentiment transformation and maintaining the original semantic content.

7. Conclusion

This overview paper presented the AraSentEval shared task, which brought forward a set of challenges addressing sentiment classification for multiple Arabic dialects as well as the generative challenge of sentiment style transfer. By benchmarking on challenging tasks such as dialectal understanding and context-preserving sentiment swap, we aim to encourage future advances in both the analytical and generative capabilities of Arabic models. We hope that AraSentEval will be a basis for future research and encourage the development of more dialect-aware and semantically consistent Arabic language models.

8. Bibliographical References

Fahad J Abdu, Raed Mughaus, Shadi Abudalfa, Moataz Ahmed, and Ahmed Abdelali. 2025. An empirical evaluation of arabic text formality transfer: a comparative study. *Language Resources and Evaluation*, 59(4):4093–4153.

Shadi I Abudalfa, Fahad J Abdu, and Maad M Alowaifeer. 2024. Arabic text formality modification: A review and future research directions. *IEEE Access*, 12:185117–185148.

Maram I Alharbi, Salmane Chafik, Saad Ezzini, Ruslan Mitkov, Tharindu Ranasinghe, and Hansi Hettiarachchi. 2025a. Ahasis: Shared task on sentiment analysis for arabic dialects. In *Proceedings of the Shared Task on Sentiment Analysis for Arabic Dialects*, pages 1–6.

Maram I Alharbi, Saad Ezzini, Hansi Hettiarachchi, Tharindu Ranasinghe, and Ruslan Mitkov. 2025b. Ador: Dataset for arabic dialects in hotel reviews: A human benchmark for sentiment analysis. In *Proceedings of the First Workshop on Advancing NLP for Low-Resource Languages*, pages 187–191.

Maram I. Alharbi, Saad Ezzini, Hansi Hettiarachchi, Tharindu Ranasinghe, and Ruslan Mitkov. 2025c. Evaluating large language models on sentiment analysis in Arabic dialects. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 67–74, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Ghadah Alwakid, Taha Osman, Mahmoud El Haj, Saad Alanazi, Mamoona Humayun, and Najm Us Sama. 2022. Muldasa: Multifactor lexical sentiment analysis of social-media content in non-standard arabic social media. *Applied Sciences*, 12(8):3806.

Samhaa R. El-Beltagy, Mona El Kalamawy, and Abu Bakr Soliman. 2017. NileTMRG at SemEval-2017 task 4: Arabic sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 790–795.

Mahmoud El-Haj. 2020. Habibi–a multi dialect multi national arabic song lyrics corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1318–1326.

#	Team/Participant Name	Sentiment Style Accuracy	BLEU	ChRF
1	Yumnahamdy	0.7554	43.00	65.36
2	L3IA	0.7430	27.22	55.04
3	Astral_Fate	0.6950	20.33	45.51
4	VGU-M.Tech-AI	0.2368	30.79	54.08

Table 2: Results of the submitted systems for Subtask 2: Arabic Sentiment Swap.

Mo El-Haj. 2026. Tarab: A multi-dialect corpus of arabic lyrics and poetry. In *Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026) at the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026)*, pages 37–46, Rabat, Morocco.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Abdusalam F. Ahmad Nwesri, Nabila Almabrouk S. Shinbir, and Amani Bahlul Sharif. 2025. Sentiment analysis on Arabic dialects: A multi-dialect benchmark. In *Proceedings of the Shared Task on Sentiment Analysis for Arabic Dialects*, pages 86–91, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*, volume 30.