DialectSentEval 2026: Arabic Dialect Sentiment Analysis and Swapping Shared Task

Saad Ezzini[1,*], Shadi Abudalfa[2,3], Maram Alharbi[4], Salmane Chafik[5], Hind Alatawi[1], Mo El-Haj[4,6], Ahmed Abdelali[7], Osamah Alnahari[1], Salima Lamsiyah[8]
[1]King Fahd University of Petroleum and Minerals, KSA, [2]Onaizah Colleges, KSA,
[3]University College of Applied Sciences, Palestine, [4]Lancaster University, UK,
[5]UM6P, Morocco, [6]VinUniversity, Vietnam, [7]El Technology, Qatar,
[8]University of Luxembourg, Luxembourg
***Correspondence:** saad.ezzini@kfupm.edu.sa

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 Dialects (DialectSentEval), hosted with the Arabic Natural Language Processing Conference (ArabicNLP 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.

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 Dialects (DialectSentEval), hosted with the Arabic Natural Language Processing Conference (ArabicNLP 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. DialectSentEval 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 DialectSentEval 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 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

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

Proceedings of The Fourth Arabic Natural Language Processing Conference (ArabicNLP 2026) - Volume 2 Shared Tasks, pages 1380–1388
October 28-29, 2026 ©2026 Association for Computational Linguistics

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., 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), DialectSentEval 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 (Abudalfa et al., 2025). 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, DialectSentEval 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, DialectSentEval 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 five major dialects: Moroccan, Egyptian, Lebanese, Jordanian, and Saudi. The data was sourced from hotel reviews from the 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 recently organised shared tasks Ahasis (Alharbi et al., 2025a) and AraSentEval (Ezzini et al., 2026).

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

This subtask draws on a dataset built as an extension of MA'AKS (Mughaus et al., 2026), comprising a total of 9,647 sentence-pair instances. Each pair centers on a shared topic but conveys contrasting sentiment orientations, with one member expressing a positive stance and the other a negative one. Of the full collection, 5,331 instances (55.3%) carry a Negative label, while the remaining 4,316 (44.7%) are tagged Positive.

Whereas the original MA'AKS resource was built predominantly around MSA, the present corpus is drawn entirely from regional Arabic dialects and spans four principal dialectal families: Gulf (5,893 pairs, 61.1% of the total), Levantine (1,655 pairs, 17.2%), Egyptian (1,561 pairs, 16.2%), and Maghrebi (538 pairs, 5.6%).

To ensure reliability, every pair passed through a rigorous validation process assessing fluency, semantic fidelity, and the correctness of the sentiment reversal. The corpus comes with fixed training, validation, and test partitions, requiring models to capture both the meaning and the sentiment structure of each sentence. These partitions consist of 6,752 pairs for training, 964 for validation, and 1,931 for testing.

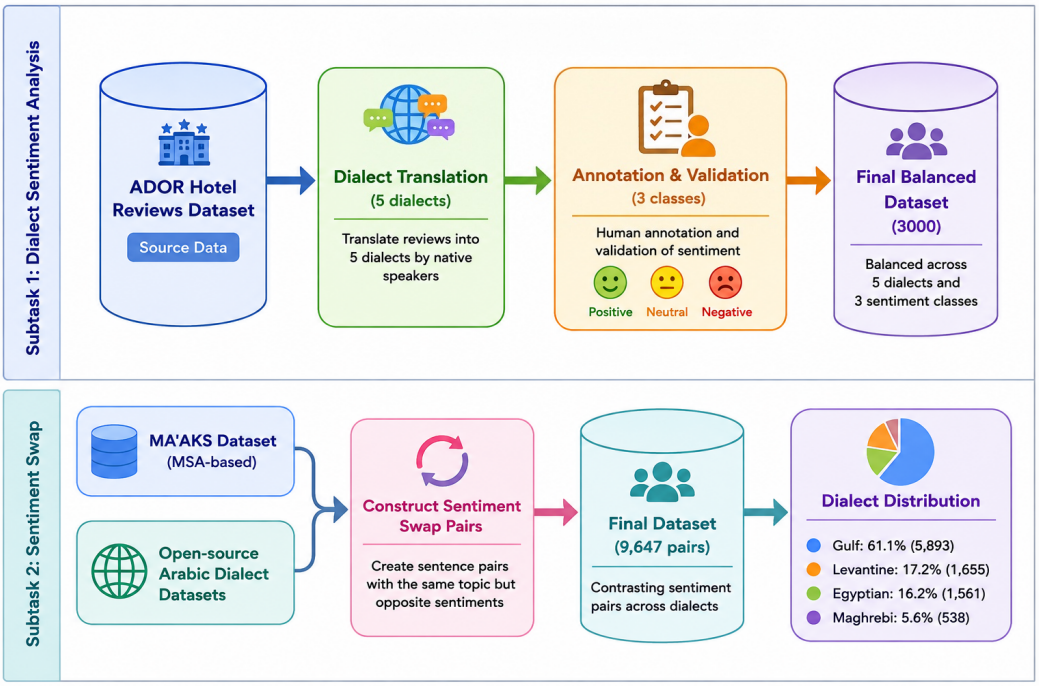

Figure 1: Data creation process for both subtasks.

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 five different and linguistically challenging Arabic dialects: Moroccan, Egyptian, Lebanese, 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, Lebanese, 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 one of several Arabic dialect groups: Maghreb, Egyptian, Gulf, and Levantine. 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 pretrained sentiment classifier from CamelLab[2] 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

Across both subtasks, a total of 15 unique teams participated in the competition, with 16 system description papers submitted. Below, we describe the methodologies and system architectures proposed by the participating teams, categorized by subtask.

5.1 Subtask 1: Arabic Dialect Sentiment Analysis

Participating teams for Subtask 1 investigated a diverse range of pretrained transformer architectures, ensembling strategies, loss function formulations, and domain-adaptation techniques to tackle multi-dialect variation and zero-shot dialect transfer.

- **ALEXIS** (Username: *momen_ibrahim-870387*) (Ibrahim et al., 2026a): Ranked **1st** overall with a Macro F1-score of **0.8667**. The team proposed an adversarial training and iterative self-training pipeline built on top of MARBERTv2. Specifically, they utilized the Fast Gradient Method (FGM) during fine-tuning to regularize the feature space against dialectal perturbations. To address zero-shot generalization on unlabelled dialects (notably Lebanese), they implemented a confidence-thresholded pseudo-labeling strategy that iteratively incorporated high-confidence test predictions back into training.
- **Axiom** (Username: *ajwad_hossain-867052*) (Hossain, 2026): Ranked **2nd** with a Macro F1-score of **0.8210**. They developed a dialect-aware ensemble framework combining MARBERTv2 and AraBERTv2. Their system utilized Stratified 5-Fold Cross-Validation, text normalization tailored to dialectal orthography, and an automated soft-voting probability aggregation scheme.
- **PTUK-NLP** (Username: *tasneemduridi-857223*) (Duridi et al., 2026): Ranked **3rd** with a Macro F1-score of **0.8174**. The team fine-tuned MARBERT with domain-specific cleaning pipelines. They demonstrated that representations learned from geographically and linguistically proximate Levantine variants (Jordanian) transfer effectively in a zero-shot fashion to Lebanese. They also provided an extensive qualitative error analysis identifying mixed-polarity reviews as the primary source of classification errors.
- **NU-SentEval** (Username: *habiba_ayman-867381*) (Ayman et al., 2026): Ranked **4th** with a Macro F1-score of **0.8140**. They developed a stacked model blending predictions from MARBERTv2, AraBERTv02, and QARiB. Crucially, they integrated Focal Loss during training to mitigate class imbalance and focus gradient updates on hard-to-classify neutral examples.
- **Gradient Descenders** (Username: *normalman159-857755*) (Nhan et al., 2026): Ranked **5th** with a Macro F1-score of **0.8010**. The team performed an empirical comparison across multiple Arabic encoders (MARBERT, ARBERT, and AraBERT), concluding that fine-tuned MARBERTv2 combined with class-weighted cross-entropy loss and linear learning-rate warm-up yielded optimal multi-dialect performance.

[2] CAMeL-Lab/bert-base-arabic-camelbert-da-sentiment

- **LahjaSensei** (Username: *ahmadalmoustafa-856400*) (Almoustafa, 2026): Ranked **6th** with a Macro F1-score of **0.7977**. They utilized fine-tuned transformer backbones accompanied by dialect-specific token normalization and prompt-based tuning.

- **NAMAA Community** (Username: *mona_-871069 / hsnhb7*) (Abdelazim et al., 2026): Ranked **7th** with a Macro F1-score of **0.7953**. They constructed an ensemble merging MARBERTv2, AraBERTv2, and AraModernBERT, using dialect classification heads to dynamically weigh base-model outputs.

- **faisalm93** (Username: *faisalm93-869067*) (Adam, 2026): Ranked **8th** with a Macro F1-score of **0.7939**. The author developed a 5-fold cross-validated MARBERTv2 bagging ensemble utilizing soft voting across fold models, Cosine Annealing learning rate schedules, and layer-wise learning rate decay.

- **Lahja** (Username: *shadenalzeer-867546*) (Alzeer, 2026): Ranked **10th** with a Macro F1-score of **0.7894**. They implemented a multi-seed ensembled MARBERT architecture designed to stabilize zero-shot transfer for the unrepresented Lebanese dialect by exploiting cross-dialect semantic regularities.

- **Northwestern** (Username: *monmon2000*) (Shlkamy and Zaghouani, 2026): Explored a comparative study of MARBERTv2 and AraBERT under various dialectal splits, demonstrating that dialect-specific pretraining provides significant robustness gains over general MSA models.

5.2 Subtask 2: Arabic Sentiment Swap

Subtask 2 saw participating systems leverage generative large language models (LLMs), parameter-efficient adapters, candidate reranking pipelines, and preference alignment techniques to tackle the complex dual objective of sentiment polarity inversion and semantic fidelity.

- **ALEXIS** (Username: *momen_ibrahim-870527*) (Ibrahim et al., 2026b): Ranked **1st** across all evaluation metrics, achieving a Sentiment Style Accuracy of **0.9581**, a BLEU score of **56.91**, and a chrF of **72.18**. They introduced a multi-stage framework combining: (1) dialect-aware Retrieval-Augmented Generation (RAG) using semantic similarity over training pairs, (2) Direct Preference Optimization (DPO) to align open-weight generative backbones towards clean polarity transfer, and (3) a classifier-guided candidate reranker that filters out candidates failing the target sentiment direction.

- **AKARROUCH** (Username: *akarrouch-871040*) (Akarrouch, 2026): Ranked **2nd** with a Sentiment Style Accuracy of **0.7851**, BLEU of **49.82**, and chrF of **70.12**. The system employed a generate-and-rerank architecture where an instruction-tuned LLM produced multiple candidate rewrites via temperature sampling, followed by a fine-tuned sentiment classifier to select the candidate maximizing target polarity probability.

- **LahjaSensei** (Username: *ahmadalmoustafa-865064*) (Almoustafa, 2026): Ranked **3rd** with a Sentiment Style Accuracy of **0.7789**, BLEU of **41.48**, and chrF of **64.45**. They proposed few-shot in-context prompting combined with Model Merging via Spherical Linear Interpolation (SLERP) across fine-tuned dialect checkpoints.

- **NAMAA** (Username: *mona_-871054 / hsnhb7-868552*) (Barmandah et al., 2026): Ranked **4th** with a Sentiment Style Accuracy of **0.7597**, BLEU of **51.58**, and chrF of **71.02**. The team introduced a cross-backbone ensembling framework combining candidate hypotheses from Qwen2.5, Llama-3, and AraT5, reranked by an auxiliary sentiment verification classifier.

- **Sabaa** (Username: *mohamedsabaa-855812*) (Sabaa and Said, 2026): Ranked **6th** with a Sentiment Style Accuracy of **0.7447**, BLEU of **49.07**, and chrF of **69.33**. They utilized fine-tuned AraT5 sequence-to-sequence models combined with an external MARBERT-based sentiment classifier to rerank beam-search hypotheses.

- **NU-SentEval** (Username: *areejkhaled-868894*) (Tamer et al., 2026): Ranked **8th** with a Sentiment Style Accuracy of **0.6898**, achieving a top-tier chrF of **72.61** and BLEU of **54.63**. They fine-tuned AraT5-base models conditioned on explicit dialect and polarity transformation prefix tokens.

- **NileNLP** (Username: *mohammed_bahgat-863352 / ehabashraf-863355*) (Bahgat et al., 2026): Ranked **9th–10th** with a Sentiment Style Accuracy of **0.6732**, BLEU of **41.38**, and chrF of **65.84**. They developed *SWAPPER*, a parameter-efficient adapter framework applying Low-Rank Adaptation (LoRA) to generative Arabic models.

6 Results and Analysis

6.1 Subtask 1: Arabic Dialect Sentiment Analysis

Table 1 reports the official evaluation results on the test set for Subtask 1, ranked by Macro F1-score.

Analysis of Subtask 1 Results. Several key findings emerge from the experimental outcomes of Subtask 1:

1. **Pretrained Dialectal Encoders Outperform Generative LLMs**: All top-performing submissions relied on dedicated dialectal encoder models—most notably MARBERTv2 and AraBERTv02. Models exposed to substantial dialectal social media data during pretraining consistently outperformed general Modern Standard Arabic (MSA) baselines and zero-shot prompting of large generative models.

2. **Robustness via Adversarial Training and Pseudo-Labeling**: The winning system, **ALEXIS**, achieved a substantial lead (+4.57% Macro F1 over 2nd place). This gain was primarily driven by Fast Gradient Method (FGM) adversarial training, which forces the model representations to remain invariant to dialectal surface lexical variations, paired with iterative self-training that effectively bootstrapped representations for the unlabelled Lebanese test instances.

3. **Ensembling and Loss Engineering**: Teams that implemented stratified K-fold ensembling (**Axiom**, **faisalm93**) and model stacking across distinct pretraining families (**NU-SentEval**, **NAMAA**) demonstrated higher stability. Furthermore, incorporating Focal Loss (**NU-SentEval**) effectively addressed the subtle semantic boundaries between neutral and mildly opinionated dialectal phrases.

4. **Zero-Shot Levantine Transferability**: Analysis by **PTUK-NLP** and **Lahja** showed that models trained on geographically proximate dialects (Jordanian) exhibit strong zero-shot cross-lingual transfer to Lebanese, confirming shared lexical roots within the Levantine dialect family.

6.2 Subtask 2: Arabic Sentiment Swap

Table 2 presents the official evaluation results for Subtask 2, sorted by the primary metric, Sentiment Style Accuracy, alongside semantic preservation metrics (BLEU and chrF).

Analysis of Subtask 2 Results. Evaluating generative sentiment swapping across Arabic dialects revealed several critical insights:

1. **The Polarity Transfer vs. Content Preservation Trade-off**: Generative sentiment style transfer poses an intrinsic tension between modifying sentiment markers (necessary for high Sentiment Style Accuracy) and retaining non-sentiment lexical content (measured by BLEU and chrF). Submissions that modified too few words achieved high BLEU but failed to invert sentiment (e.g., lower accuracy), while unconstrained generative rewrites suffered severe lexical drift.

2. **The Impact of Preference Optimization (DPO)**: The winning team, **ALEXIS**, dominated across all dimensions (0.9581 accuracy, 56.91 BLEU, 72.18 chrF). By directly training on paired contrastive preferences using DPO, their system learned to apply minimal, surgical edits to dialectal sentiment words (e.g., negating adjectives or swapping affective idioms) while strictly preserving sentence structure.

3. **Generate-and-Rerank Pipelines as an Essential Safety Mechanism**: Teams employing generate-and-rerank pipelines (**AKARROUCH**, **NAMAA**, **Sabaa**) achieved consistently higher style transfer accuracy ($> 74\%$) than unguided generation. Filtering candidate rewrites with an external sentiment classifier prevented output hallucination and ensured the target polarity was reliably achieved.

4. **Model Merging and PEFT for Dialect Specialization**: Advanced parameter-efficient techniques, such as SLERP model merging across dialect checkpoints (**LahjaSensei**) and

#	Team Name	CodaBench Username	Macro F1
1	**ALEXIS**	*momen_ibrahim-870387*	**0.8667**
2	**Axiom**	*ajwad_hossain-867052*	0.8210
3	**PTUK-NLP**	*tasneemduridi-857223*	0.8174
4	**NU-SentEval**	*habiba_ayman-867381*	0.8140
5	**Gradient Descenders**	*normalman159-857755*	0.8010
6	**LahjaSensei**	*ahmadalmoustafa-856400*	0.7977
7	**NAMAA Community**	*mona_-871069*	0.7953
8	**faisalm93**	*faisalm93-869067*	0.7939
9	–	*hoanguitcs-866914*	0.7936
10	**Lahja**	*shadenalzeer-867546*	0.7894
11	–	*hh03112005-869469*	0.7859
12	–	*23020073-870754*	0.7722

Table 1: Official evaluation results for Subtask 1: Arabic Dialect Sentiment Analysis.

#	Team Name	CodaBench Username	Sentiment Style Accuracy	BLEU	chrF
1	**ALEXIS**	*momen_ibrahim-870527*	**0.9581**	**56.91**	72.18
2	**AKARROUCH**	*akarrouch-871040*	0.7851	49.82	70.12
3	**LahjaSensei**	*ahmadalmoustafa-865064*	0.7789	41.48	64.45
4	**NAMAA**	*mona_-871054*	0.7597	51.58	71.02
5	–	*melmoussaoui-870973*	0.7530	49.03	69.53
6	**Sabaa**	*mohamedsabaa-855812*	0.7447	49.07	69.33
7	**NAMAA**	*hsnhb7-868552*	0.7343	46.80	67.10
8	**NU-SentEval**	*areejkhaled-868894*	0.6898	54.63	**72.61**
9	**NileNLP**	*mohammed_bahgat-863352*	0.6732	41.38	65.84
10	**NileNLP**	*ehabashraf-863355*	0.6732	41.38	65.84
11	–	*ahmad_alshattha-870972*	0.6354	25.07	56.33
12	–	*hoanguitcs-868178*	0.6220	49.27	68.26
13	**Gradient Descenders**	*normalman159-868503*	0.5997	11.31	38.22
14	–	*mustafaelkady-870032*	0.4772	0.03	10.04
15	–	*mohamedmhaouach-872442*	0.2807	43.06	64.61

Table 2: Official evaluation results for Subtask 2: Arabic Sentiment Swap.

Low-Rank Adaptation (**NileNLP**), proved competitive while requiring substantially lower computational footprints during fine-tuning.

7 Conclusion

This overview paper presented the results and findings of the DialectSentEval shared task, held at ArabicNLP 2026. The shared task benchmarked two core dimensions of dialectal NLP: multi-class Arabic dialect sentiment classification (Subtask 1) and generative sentiment style transfer across four major Arabic dialect groups (Subtask 2).

A total of 15 teams actively participated in the competition. The evaluation demonstrated the continued superiority of dialect-pretrained masked encoders when paired with adversarial regularization and self-training for classification, as well as the immense potential of Direct Preference Optimization (DPO) and classifier-guided reranking for controlled generative style transfer in low-resource and dialectally diverse environments.

References

Mona Abdelazim, Hassan Barmandah, Yousra Elghawi, and Omer Nacar. 2026. NAMAA community at DialectSentEval 2026 subtask 1: Dialect-aware arabic sentiment classification with pretrained transformer models. In *Proceedings of the Fourth Arabic Natural Language Processing Conference (ArabicNLP 2026)*.

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 Abudalfa, Fahad Abdu, and Saad Ezzini. 2025. Emerging trends and challenges in arabic text style transfer: A comprehensive survey. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

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.

Faisal Muhammad Adam. 2026. faisalm93 at DialectSentEval 2026: A five-fold MARBERTv2 ensemble for arabic dialect sentiment analysis. In *Proceedings*

of the Fourth Arabic Natural Language Processing Conference (ArabicNLP 2026).

Mohamed Akarrouch. 2026. AKARROUCH at DialectSentEval 2026: A generate-and-rerank approach for arabic dialect sentiment swapping. In *Proceedings of the Fourth Arabic Natural Language Processing Conference (ArabicNLP 2026)*.

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.

Ahmad Almoustafa. 2026. LahjaSensei at DialectSentEval 2026: Few-shot prompting and model merging for sentiment swap. In *Proceedings of the Fourth Arabic Natural Language Processing Conference (ArabicNLP 2026)*.

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 nonstandard arabic social media. *Applied Sciences*, 12(8):3806.

Shaden Alzeer. 2026. Lahja at DialectSentEval shared task subtask-1: Seed-ensembled MARBERT for multi-dialect arabic sentiment analysis with zero-shot lebanese transfer. In *Proceedings of the Fourth Arabic Natural Language Processing Conference (ArabicNLP 2026)*.

Habiba Ayman, Sara Mohamed, Esraa Mohamed, Alaa Tamer, and Areej Khaled. 2026. NU-SentEval at DialectSentEval shared task subtask-1: Stacking pretrained language models with focal loss for arabic dialect sentiment analysis. In *Proceedings of the Fourth Arabic Natural Language Processing Conference (ArabicNLP 2026)*.

Mohamed Bahgat, Mahmoud Alshrief, Ehab Ashraf, and Abdelgwad Elashry. 2026. NileNLP at DialectSentEval 2026: SWAPPER — sentiment SWAPping with parameter-efficient adapters for arabic dialects. In *Proceedings of the Fourth Arabic Natural Language Processing Conference (ArabicNLP 2026)*.

Hassan Barmandah, Yousra Elghawi, Mona Abdelazim, and Omer Nacar. 2026. NAMAA at DialectSentEval 2026 subtask 2: From reranked sampling to cross-backbone ensembling for arabic dialectal sentiment swap. In *Proceedings of the Fourth Arabic Natural Language Processing Conference (ArabicNLP 2026)*.

Tasneem Duridi, Asma Murad Mallak, Balsam Khaleel, Elaa Jaber, and Areej Jaber. 2026. PTUK-NLP at DialectSentEval shared task subtask-1: A simple MARBERT baseline, the missing lebanese dialect, and an error analysis of mixed-sentiment reviews. In *Proceedings of the Fourth Arabic Natural Language Processing Conference (ArabicNLP 2026)*.

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.

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.

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.

Md. Ajwad Hossain. 2026. Axiom at DialectSentEval 2026: Dialect-aware ensembling of MARBERTv2 and AraBERTv2 for arabic sentiment analysis. In *Proceedings of the Fourth Arabic Natural Language Processing Conference (ArabicNLP 2026)*.

Momen Ibrahim, Mariam E. Hassib, Ahmed Salem, and Marwan Torki. 2026a. ALEXIS at DialectSentEval subtask 1: Adversarial training and iterative self-training for zero-shot dialect sentiment analysis. In *Proceedings of the Fourth Arabic Natural Language Processing Conference (ArabicNLP 2026)*.

Momen Ibrahim, Mariam E. Hassib, Ahmed Salem, and Marwan Torki. 2026b. ALEXIS at DialectSentEval subtask 2: Retrieval-augmented, preference-tuned generation for arabic sentiment swap. In *Proceedings of the Fourth Arabic Natural Language Processing Conference (ArabicNLP 2026)*.

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.

Tran Phuoc Thanh Nhan, Le Thanh Trong, and Dang Van Thin. 2026. Gradient descenders at DialectSentEval 2026: A comparative study of pre-trained arabic transformers for multi-dialect sentiment analysis. In *Proceedings of the Fourth Arabic Natural Language Processing Conference (ArabicNLP 2026)*.

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.

Mohamed Sabaa and Amr Said. 2026. Sabaa at DialectSentEval shared task subtask-2: Classifier-guided candidate reranking for arabic dialectal sentiment swap. In *Proceedings of the Fourth Arabic Natural Language Processing Conference (ArabicNLP 2026)*.

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.

Eman Sedqy Shlkamy and Wajdi Zaghouani. 2026. Northwestern at DialectSentEval 2026: A comparative study of MARBERTv2 and AraBERT for arabic dialect sentiment analysis. In *Proceedings of the Fourth Arabic Natural Language Processing Conference (ArabicNLP 2026)*.

Alaa Tamer, Areej Khaled, and Mariam Mazen. 2026. NU-SentEval at DialectSentEval shared task subtask-2: Fine-tuning AraT5 for arabic dialect sentiment swap. In *Proceedings of the Fourth Arabic Natural Language Processing Conference (ArabicNLP 2026)*.