Codezone Research Group at AraSentEval Shared Task: Arabic Sentiment Swap Beyond Negation Prepending, Benchmarking Multilingual T5 Against Large Language Models on the MA'AKS Corpus

Abdulkadir Shehu Bichi[1], **Sarah Yassine**[2]

[1]Vivekananda Global University Jaipur, [2]Lebanese University/Lebanon

Correspondence: 24wtec3csml001@vgu.ac.in, abdulkadir.bichi@babaahmeduniversity.edu.ng,

Abstract

We launched ASBN-MT5, the system for Arabic Sentiment Swap, which performs the task of inverting the sentiment of a sentence while keeping the meaning intact. This is a sequence-to-sequence task. We demonstrate *ASBN-MT5: mT5*, which is a MultiLingual T5 model, fine-tuned on the provided dataset of the AraSentEval 2026 Shared Task. We describe the data as the first of its kind for the Arabic language, as *MAAKS* is the first manually composed, parallel, cross-linguistic corpus for the Arabic language. With the preliminary results of *Sentiment Flip* for the task of Sentiment Inversion, we have recorded a rate of 59.5% for positive to negative conversions and 58.5% for negative to positive conversions, while maintaining an average similarity to the original sentences of 0.955. We present the *Arabic prompts* and a neuro-developmental (Deep Learning) recipe. Due to the evaluation criteria which include Exact Match, Flip Success, Surface Similarity, and Quality of Output, we restrict the use of Prepended Negation as the main technique and recommend the use of LLMs designed for the Arabic language in the near future.

1 Introduction

Sentiment analysis has focused primarily on *detecting* polarity; a more difficult problem is *inverting* polarity while keeping the rest of the message the same. This task is sometimes called *sentiment swap* or controlled sentiment transfer (Jin et al., 2022). In the case of Arabic, this problem is made more difficult due to the richness of its morphology and its dialectal variation. In Arabic, negation can be expressed using proclitic particles such as *lā*, *lam*, and *lan*, as well as through word-order shifts and the use of different lexemes (Farha and Magdy, 2024). In the last couple of years, the emergence of Arabic-centered large language models (LLMs), such as Jais (Sengupta et al., 2023) and AceGPT (Huang et al., 2024), has changed this. However, the supervised fine-tuning of multilingual seq2seq models is still robust, and it is the approach of choice in low-resource generation. The **AraSentEval 2026 Shared Task** (Subtask 2) (Ezzini et al., 2026) sets the first evaluation benchmark for Arabic sentiment swap, based on the MA'AKS corpus (Mughaus et al., 2026). We participated with **ASBN-MT5**, a fine-tuned mT5-small (Xue et al., 2021) system that generates sentiment-inverted Arabic text based on a given source sentence and a polarity label attached to it.

Our contributions are (i) an **Arabic-script prompt strategy** where the prompt to the encoder is written in Arabic script and encodes the desired target polarity; (ii) a **memory-efficient training recipe** that fits on a single NVIDIA T4; and (iii) a **multi-faceted evaluation** that examines exact match, flip detection heuristics, surface similarity, and various quality metrics.

2 Related Work

Text Style Transfer. The early methods of style transfer consisted of disentangled latent-space methods (Shen et al., 2017) and back-translation (Prabhumoye et al., 2018), and more recently have been shifting towards LLM-based techniques. Mukherjee et al. (2024) evaluates the performance of LLMs on sentiment transfer in a cross-lingual context and asserts that the performance of a model that has been fine-tuned is always better than zero- and few-shot prompting, which serves as a direct motivation for our supervised approach. The use of diffusion models has also been explored for more detailed style transfer (Lyu et al., 2023), including the augmentation of Arabic mental health texts (Mankarious and Zirikly, 2025).

Arabic Sentiment Analysis (2023–2026). With the advent of pre-trained transformers, the field of Arabic NLP has made great strides. Almaqtari et al. (2024) show that AraBERT-based CNN+BiGRU hybrids produce notable results on consumer reviews. Aljomah et al. (2025) show that CAMeLBERT, which has been fine-tuned on 56 thousand samples of MARSA, performs exceptionally well on the noisy dialectal text. In Zouidine and Khalil (2025), LLaMA, Mixtral, and Gemma are evaluated under zero- and few-shot conditions, where task-specific transformer models are found to outperform general LLMs for Arabic sentiment classification.

Arabic LLMs for Generation (2023–2025). A new milestone for developing Arabic LLMs is the release of Jais (Sengupta et al., 2023), a 13B parameter Arabic-English LLM. Huang et al. (2024) Jais was enriched with instruction tuning via RLHF, and ArabianGPT (Koubaa et al., 2025) presented *AraNizer*, a native Arabic tokenizer that attained 95% accuracy after senti-

Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks, pages 288–291

ment fine-tuning. These models have been assessed on the Arabic formality transfer and dialect-to-MSA translation tasks (Abdu et al., 2025), where the benefits of Arabic-centric pretraining have been consistently recorded over the multilingual baselines.

Arabic Sentiment Swap Resources. MA'AKS (Mughaus et al., 2026) is the first sentiment-style transfer parallel corpus of Arabic and it is a 5k MSA sentence pair. In the benchmarking of AceGPT, Jais and Llama-3 on MA'AKS, it demonstrates that fine-tuning outruns zero-shot and few-shot settings by a large margin, and it is Jais that has the greatest advantage in few-shot settings. The 2026 shared task of AraSentEval (Ezzini et al., 2026) extends this resource to multi-domain and *system* comparative evaluation for the first time.

3 System Description

3.1 Task Formulation

Given a source sentence x with polarity $p \in \{Positive, Negative\}$, the model generates y with inverted polarity $\bar{p}$: $y = f(x, \bar{p})$.

3.2 Model and Prompt Strategy

We fine-tune **mT5-small** (300 M parameters) (Xue et al., 2021), pre-trained on the mC4 corpus, including substantial Arabic web text. The encoder receives an Arabic-script prompt x:

where (negative) or (positive) is selected based on the desired output polarity. Using Arabic script for the conditioning token rather than an English word allows the encoder to leverage pre-trained Arabic subword representations.

3.3 Training Configuration

This section lists all settings. Gradient accumulation achieves an effective batch of 16; gradient checkpointing reduces peak memory from ∼4.5 GB to ∼2.2 GB. At inference, beam search($b = 4$) with a no-repeat bigram penalty promotes output diversity. It consists of hyperparameter values such as Backbone: mT5-small (300 M params), Per-device batch: 2, Gradient accum: 8 steps (eff. batch = 16) Learning rate: 5×10^{-5}, AdamW, Weight decay: 0.01, LR schedule: Linear warm-up (200 steps), Gradient clipping: 1.0, Max seq. length: 96 tokens, Epochs: 5, Inference: Beam (b=4), no-rep 2-gram, and Hardware: NVIDIA Tesla T4 (16 GB)

4 Experiments

4.1 Dataset

The AraSentEval 2026 Subtask 2 corpus (Ezzini et al., 2026), based on MA'AKS (Mughaus et al., 2026), contains 7,578 annotated pairs: 6,263 training, 1,315 validation, and 646 test. Sources span product reviews, book critiques, and hospitality feedback in Egyptian and Gulf dialects and MSA (mean source length 11.0 words).

4.2 Evaluation Metrics

We report: **Exact Match (EM):** character-level identity against the human reference; **Sentiment Flip Rate (SFR):** a heuristic checking whether Arabic negation markers appear or disappear in the expected direction (changed outputs with ambiguous polarity receive the benefit of the doubt); **Text Similarity:** Sequence Matcher character ratio; and **Quality Indicators:** empty-prediction and source-copy rates. For shared task comparison, organizers additionally report **Sentiment Style Accuracy (SSA)**, **BLEU**, and **chrF**.

4.3 Validation Performance

Table 1 reports validation metrics. SFR reaches ≈59 % for both directions with zero empty predictions. The 37.6 % source-copy rate is the dominant failure mode. Figure **??** shows the full evaluation dashboard; the text similarity distribution is strongly right-skewed (median 0.977), confirming conservative, small-edit generations.

Metric	Pos→Neg	Neg→Pos
Sentiment Flip Rate (%)	59.46	58.48
Exact Match (%)	1.53	0.80
Avg. Text Similarity	0.955	0.954
Empty Predictions (%)	0.00	0.00
Overall		
Identical to Source (%)	37.64	
Avg. Pred. Length (words)	10.3	

Table 1: Validation evaluation results for ASBN-MT5 (1,315 samples).

4.4 Comparison of Leaderboards from Shared Tasks

The table 2 contains the Comparative Sentiment Analysis evaluation for AraSentEval 2026 Subtask 2, for all benchmarked systems, presented by the organizers as ASE, BLEU, and chrF, where SSA = Sentiment Style Accuracy; chrF = character n-gram F-score. Our submission is highlighted in *italic*. With Rank 4, asbichi362 (our proposed system) reached an SSA of 0.237, with BLEU = 30.79 and chrF = 54.08. The highest-ranked system, yumnahamdy, reaches SSA = 0.755 and chrF = 65.36, which points out that the system is very high on sentiment accuracy and surface fidelity. It is also important to mention that ASBN-MT5 achieved a competitive BLEU of 30.79, which is better than Rank 2, which achieved 27.22, and Rank 3, with 20.33, which means that while our mT5-small has fewer issues than other models at inverting or flipping the sentiment and producing a correct bottom-line result, the outputs on the surface level are also very close to the bottom-line result.

The difference between our system and the top systems in SSA shows the inadequacy, which is known, of the negation-particle prepending for the polarity inversion strategy. The generated output is very similar to the

input/output and often does not pass the sentiment classifier that is used to measure the SSA. Our expectations are that new models for processing and understanding new levels of branching subtexts in the Arabic language will reduce the gap by more than 95 percent.

Rank	Team	SSA	BLEU	chrF
1	yumnahamdy	0.7554	43.00	65.36
2	beatchalvador	0.7430	27.22	55.04
3	astral_fate	0.6950	20.33	45.51
4	asbichi362 (ours)	0.2368	30.79	54.08

Table 2: Official AraSentEval 2026 Subtask 2 leaderboard (test set)

4.5 Qualitative Analysis

Three-generation patterns emerge. **(i) Negation prepending**: the dominant strategy is prepending to affirmative-verb sentences (e.g., ...); this fails when sentiment is lexically encoded. **(ii) Lexical substitution**: a smaller fraction replaces sentiment-bearing adjectives, producing outputs closer to human references. **(iii) Source copying**: 37.6 % of outputs are identical to the source. Comparing with Mughaus et al. (2026), where fine-tuned Llama-3 achieves stronger lexical substitution, suggests that Arabic-centric LLMs offer the clearest path to higher-quality sentiment inversion.

5 Discussion and Conclusion

ASBN-MT5 is a notable case that demonstrates how to operate Arabic sentiment inversion with relatively small computational cost (55 min on a single T4). Its SFR over the negations is notable, and zero empty prediction instances attest to ASBN-MT5's stable generation capabilities. In Table 2, ASBN-MT5 is ranked 4th for SSA. This is contextualized with a BLEU score of 30.79, which is greater than the 2nd- and 3rd-ranked ASBN-MT5 systems, demonstrating that, along with being lower in rank for the SSA, ASBN-MT5 shows surface-level fidelity higher than those systems. There are still a number of limitations to the current work. Most notable is the 37.6% source-copy rate that affects SSA. This can be diminished to a degree with the use of a copy-penalty decoding or minimum-edit constraints. BERTScore (Zhang et al., 2020) is a metric that can be used to address the shortage of measuring in the Arabic language, in which case an EM in a character level is of little value, and that is still the case for the Arabic language. BERTScore is also a metric that can be used to address the shortage of measuring in the Arabic language. Finally, mT5-small has shown to be outscaled by Arabic-centric LLMs (Sengupta et al., 2023; Huang et al., 2024) that demonstrate higher lexical transfer for MA'AKS (Mughaus et al., 2026). Instruction tuning with RLHF (Ouyang et al., 2022) and sentiment classifiers will further enhance these systems.

We summarize our contribution to AraSentEval 2026 Subtask 2 (Ezzini et al., 2026) as **ASBN-MT5**. From the MA'AKS corpus (Mughaus et al., 2026), we report ~59 % flip success with zero empty predictions and a BLEU score of 30.79, competitive with the official test set. This offers an efficient and reproducible baseline for Arabic sentiment swap. Code and weights will be made public after acceptance.

Limitations

The negation-heuristic SFR is limited to the lexical sentiment alterations, mT5-small is smaller than the most current Arabic LLMs, a source-copy rate of 37.6 % suggests the generation is failing a lot, and the test-set targets are withheld by the organizers.

References

Fahad Abdu, Raed Mughaus, Shadi Abudalfa, Mohammed Ahmed, and Ahmed Abdelali. 2025. An empirical evaluation of Arabic text formality transfer: A comparative study. *Language Resources and Evaluation*.

Fayez Aljomah, Lama Aldhafeeri, Mohammed Alfadel, Saad Alshahrani, Qaisar Abbas, and 1 others. 2025. Enhancing Arabic sentiment analysis with pre-trained CAMeLBERT: A case study on noisy texts. *Computers, Materials & Continua*, 84(3):5317–5335.

Hamzah Almaqtari, Fang Zeng, and Abdulaziz Mohammed. 2024. Enhancing Arabic sentiment analysis of consumer reviews: Machine learning and deep learning methods based on NLP. *Algorithms*, 17(11):495.

Saad Ezzini, Paul Rayson, Shadi Abudalfa, Maram Alharbi, and Mo El-Haj. 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.

Ibrahim Abu Farha and Walid Magdy. 2024. Arabic sentiment analysis: Challenges and opportunities. *ACM Computing Surveys*.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Mosen Alharthi, Bang An, and 1 others. 2024. AceGPT, localizing large language models in Arabic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 8139–8163, Mexico City, Mexico. 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.

Anis Koubaa, Adel Ammar, Lahouari Ghouti, Omer Nacar, and Serry Sibaee. 2025. ArabianGPT: An Arabic language adaptation of GPT-2. In *Generative AI*

and Large Language Models: Opportunities, Challenges, and Applications, volume 1214 of *Studies in Computational Intelligence*. Springer, Cham.

Yuning Lyu, Tongshuang Luo, Jinliang Shi, Todd Hollon, and Honglak Lee. 2023. Fine-grained text style transfer with diffusion-based language models. In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 65–74. Association for Computational Linguistics.

Sarah Mankarious and Ayah Zirikly. 2025. Style transfer as bias mitigation: Diffusion models for synthetic mental health text for Arabic. In *arXiv preprint arXiv:2601.14124*.

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.

Sourabrata Mukherjee, Atul Kr. Ojha, Pruthwik Mujadia, and Ondrej Dusek. 2024. Are large language models actually good at text style transfer? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Bangkok, Thailand. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W. Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Neha Sengupta, Sunil Kumar Siddiqui, Basil Boregowda, and 1 others. 2023. Jais and Jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

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, pages 6830–6841.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 483–498, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Mohamed Zouidine and Mohamed Khalil. 2025. Large language models for Arabic sentiment analysis and machine translation. *Engineering, Technology & Applied Science Research*, 15(2):20737–20742.