REGLAT at SemEval-2026 Task 9: Enhancing Arabic Online Polarization Detection Using AraBERT and Synonym Replacement Augmentation

Ahmed M. Fetouh[1] Rahmath Mohammed[2] Omer Dawood[2] Mariam Labib[3] Nsrin Ashraf[3] and Hamada Nayel[1,2]

[1]Department of Computer Science, Faculty of Computers and AI, Benha University, Egypt
[2]Department of Computer Engineering and Information, College of Engineering, Prince Sattam Bin Abdulaziz University, Al-Kharj 16273, Saudi Arabia
[3]Computer Engineering, Elsewedy University of Technology, Cairo, Egypt

Correspondence: ahmed.megahed@fci.bu.edu.eg

Abstract

In this paper, we present our system, which was submitted to SemEval-2026 Task 9 (Subtask 1: Polarization Detection) and focuses on binary classification of polarized content in Arabic social media text. To address Arabic linguistic variations, we propose a single-model approach that combines fine-tuned AraBERT with synonym-based data augmentation. On the Arabic bind set, our method achieves a competitive macro F1-score of 0.831 and an accuracy of 0.833. Among the 45 participating teams, our system ranked 11th overall, with a performance gap of 0.018 macro F1 from the top-ranked team (0.8488). The results show that a fine-tuned AraBERT with synonym replacement is a strong, simple, and reproducible baseline that outperforms more complex setups in dealing with Arabic attitude polarization nuances.

1 Introduction

Arabic, spoken by over 400 million people across more than 20 countries in the Middle East and North Africa (MENA), is one of the world's most widely used languages and serves as the official language across the Arab world, as shown in Figure 1. Arabic's diverse linguistic and geographic reach has made it an important language for natural language processing (NLP) research, especially as social media platforms have emerged as central spaces for public discourse, political debate, and opinion formation in Arabic-speaking regions.

The proliferation of social media has completely changed communication in the Arab world, opening up opportunities for both dialogue and division. Empirical studies show that platforms like Twitter and Facebook amplify partisan clustering and polarization dynamics in Arabic-speaking contexts. Weber et al. (2013) used Twitter retweet networks to measure Islamist-secular polarization in Egypt, revealing distinct partisan cohesion patterns. Borge-Holthoefer et al. (2015) used network and content analysis to show that online polarization during Egypt's 2013 political events was characterized by limited cross-side engagement and significant volume shifts during protest phases. Kazkaz (2022) found that influential social media accounts contribute to political and religious polarization in Jordanian public discourse, emphasizing the growing role of digital opinion leaders in shaping polarized attitudes.

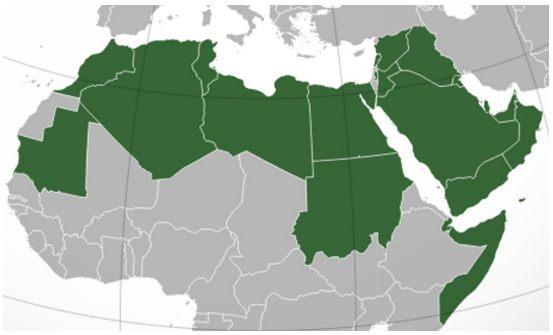

Figure 1: Political map of the Arab world

Despite the critical importance of understanding and detecting polarization in Arabic social media, Arabic natural language processing faces ongoing technical challenges that complicate sentiment analysis and polarization detection tasks. Arabic has significant dialectal variation across regions, with Modern Standard Arabic (MSA) coexisting with numerous colloquial dialects that differ significantly in vocabulary, morphology, and syntax (Hengle et al., 2021). Tokenization and feature extraction in Arabic are challenging due to its morphological complexity, which includes extensive inflection, derivation, and cliticization. Arabic suffers from a scarcity of annotated datasets for specialized tasks, including stance detection and polarization classification (AbuElAtta et al., 2023; AlRowais and Alsaeed, 2023).

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1779–1783

The remainder of the paper is organized as follows: Section 2 discusses related research on polarization detection and Arabic NLP. Section 3 outlines the task definition and dataset. Section 4 explains our methodology, including model architecture and augmentation strategy. Section 5 presents experimental results and analysis. Finally, Section 6 concludes with a discussion and future directions.

2 Background

SemEval-2026 Task 9 (Subtask 1: Polarization Detection) requires binary classification of social media posts as polarized (True) or non-polarized (False). Training and development datasets are provided for 22 languages, including Arabic . Each language has approximately 3,000-5,000 annotated instances sourced from Facebook, X, Reddit, Bluesky, and news websites. The primary metric used in evaluation is the macro F1-score. Key challenges include detecting subtle attitude polarization in noisy short texts, dealing with multilingual and cultural diversity in low-resource settings, and generalizing across events. A BERT-based baseline model has been provided by task organizers[1].

3 Data

We use the Arabic subset of the POLAR benchmark dataset introduced by POLAR: A Benchmark for Multilingual, Multicultural, and Multi-Event Online Polarization (Naseem et al., 2026b,a). The dataset is part of SemEval-2026 Task 9 (Subtask 1: Polarization Detection) and consists of binary labels indicating whether a text is polarized or non-polarized. The Arabic training split contains 3,381 annotated social media posts collected from diverse online platforms and events.

4 System Overview

As shown in Figure 2 our proposed system consists of four stages: preprocessing, augmentation, training the model and evaluation.

4.1 Preprocessing

Arabic social media texts are noisy and dialect-rich, requiring cleaning before feeding them into AraBERT (Antoun et al., 2020). We applied the official `ArabertPreprocessor` from the AraBERT library[2] to normalize and clean the raw text. The preprocessor performs the following key operations:

- Removal and normalization of diacritics (tashkeel) and tatweel (elongation).
- Standardization of Arabic letter variants (e.g., alef, yaa, and hamza forms).
- Handling of clitics, punctuation, numbers, and common social media artifacts.
- General whitespace and repetition cleanup.

This step produces standardized `cleaned_text` ready for tokenization, improving model compatibility with dialectal and informal Arabic. To illustrate the effect of preprocessing, Table 1 shows two original instances from the dataset and their cleaned versions after applying the ArabertPreprocessor.

4.2 Data Augmentation

To improve generalization and handle lexical/dialectal variation in Arabic polarized texts, we apply synonym replacement augmentation. A small hand-crafted dictionary of Arabic sentiment synonyms is used (e.g., سيء → [ردىء, ضعيف, فاشل, مخزي]) .

For each original sample:

- Generate 1–3 augmented copies (randomly selected).
- Replace 1–3 words with random synonyms from the dictionary.
- Skip if the new text is identical to the original.

This expands the dataset from 3,380 to 6,032 examples while keeping original labels.

Example:

Original_instance:-

هذا الشخص سيء جداً و غبي

Augmented instance:-

هذا الشخص ردىء جداً و أحمق

[1] https://github.com/Polar-SemEval/SemEval2026-task9/blob/main/starter.ipynb

[2] https://huggingface.co/aubmindlab/bert-base-arabertv2

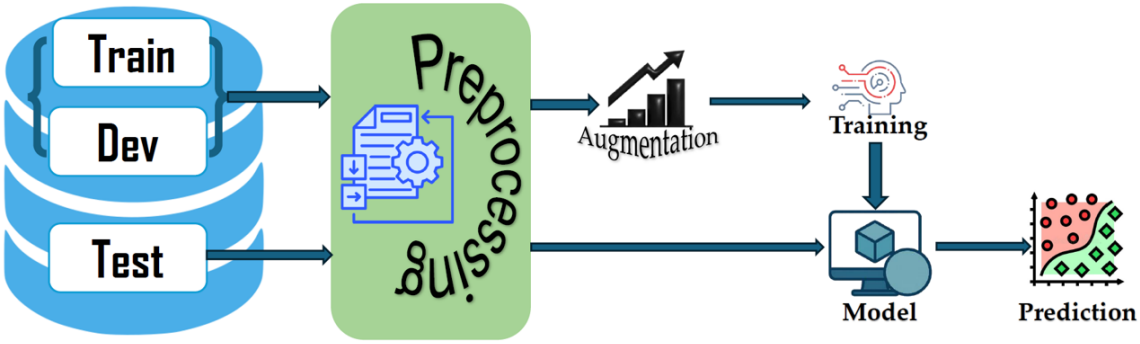

Figure 2: system architecture

Table 1: Examples of original and preprocessed Arabic social media texts.

Original Text	Preprocessed Text
عادي بعد هذه الهجمات #فان_دايك يشوتها في ارنولد وتدخل بين ارجل اليسون للشباك .... وشخصية البطل #الاتحاد_الفتح مازال فريق ابوقمطه يلعب باسلوب !!! الهجمات المرتدة حيلة الضعيف	عادي بعد هذه الهجمات فان دايك يشوتها في ارنولد وتدخل بين ارجل اليسون للشباك شخصية البطل الاتحاد الفتح مازال الفريق يلعب باسلوب الهجمات المرتدة حيلة الضعيف
أحلام رووووووووعه بكل شي احببببببببببببببببببببك من كل قلبي	أحلام روعة بكل شي احبك من كل قلبي

4.3 Model Selection

In this work, we adopted **AraBERTv2** (Antoun et al., 2020), specifically the `aubmindlab/bert-base-arabertv2` variant, as our language model. AraBERTv2 is a transformer-based model pretrained on a large Arabic corpus, optimized for MSA and robust to noisy, user-generated text. Leveraging AraBERTv2 along with its corresponding preprocessing ensures that the input text aligns with the model's pretraining assumptions, improving stability and performance during fine-tuning.

4.4 Fine-tuning

We fine-tune the model on our augmented dataset for polarization classification task. The datasets are first converted into the `Dataset` format from `Hugging Face Datasets` (Wolf et al., 2020), and tokenized using the AraBERTv2 tokenizer with a maximum sequence length of **256** tokens.

The model is trained using the `Trainer` API, with evaluation at each epoch and early model selection based on the macro F1-score. Accuracy and macro F1 are used as evaluation metrics. Training is performed for 6 epochs with a learning rate of 2×10^{-5}, batch size of 32 for training, and batch size of 16 for evaluation.

4.5 Evaluation

The performance is measured using the **macro F1-score**, computed over the predicted labels and the corresponding gold labels. This metric is chosen to equally weight all classes, including minority ones, which is especially important for imbalanced datasets.

4.6 Experimental Results

AraBERTv2 fine tuned on our augmented Arabic polarization dataset and evaluated for its performance on the validation set. Table 2 shows the training and validation metrics per epoch.

Table 2: Training and validation scores over epochs.

Epoch	Training Loss	Validation Loss	Accuracy	Macro F1
1	0.853	0.301	0.884	0.884
2	0.601	0.250	0.905	0.904
3	0.384	0.245	0.924	0.924
4	0.241	0.273	0.933	0.932
5	0.241	0.302	0.939	0.939
6	0.241	0.304	0.936	0.936

Additionally, the confusion matrix in Figure 3 confirms the model's robust classification performance for both classes.

5 Results

We evaluated our proposed system on the Arabic portion of Subtask 1 (Polarization Detection). The

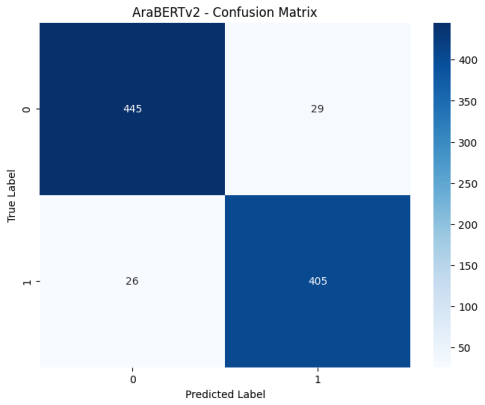

Figure 3: Confusion matrix for AraBERTv2 on the validation set.

official primary metric is macro F1. Table 3 reports the binary classification performance on the development and blind test sets. On the development set, the model obtains a macro F1 of 0.790 and an accuracy of 0.793. Performance improves notably on the blind test set, reaching a macro F1 of 0.831 and an accuracy of 0.833.

The 4% improvement in macro F1 on the blind test set demonstrates effective generalization due to synonym replacement augmentation addressing data scarcity and dialectal/lexical variations that are common in Arabic polarized social media content.

5.1 Error Analysis

Error analysis has been conducted to show the limitations of the proposed model. Misclassifications frequently occur in dialectal texts, where informal spelling and region-specific vocabulary reduce model effectiveness compared to MSA. Additionally, the model struggles with implicitly polarized content, sarcasm, and context-dependent expressions that lack explicit sentiment cues. False positives are often associated with emotionally intense but non-polarized statements, while false negatives arise from subtle or indirect polarization. Although synonym-based augmentation improves generalization, it occasionally introduces semantic drift, leading to noisy training instances. These findings highlight the impact of dialect-aware modeling and context-sensitive representations.

Our single-model approach offers a strong, reproducible, and computationally efficient solution. Future extensions could include cross-lingual fine-tuning or ensemble methods to push performance further in multilingual scenarios.

6 Conclusion

In this paper, we present a simple effective system for Arabic Polarization Detection. By fine-tuning AraBERT and applying synonym-based data augmentation, we addressed key challenges of dialectal variation, data scarcity in detecting attitude polarization. Our approach achieved a competitive macro F1-score of 0.831 and accuracy of 0.833 on the Arabic blind test set, demonstrating strong generalization and outperforming the task's basic baseline. The results highlight that a single, well-tuned transformer model combined with lightweight lexical augmentation can serve as a robust, reproducible baseline for low-resource Arabic NLP tasks like polarization detection. Future work could explore cross-lingual transfer learning, integration of dialect-specific embeddings, or ensemble methods to further improve performance across the multilingual setting. Overall, our method shows the value of targeted augmentation strategies in enhancing transformer-based models for nuanced social media analysis in diverse linguistic contexts.

References

Ahmed H. AbuElAtta, Mahmoud Sobhy, Ahmed A. El-Sawy, and Hamada Nayel. 2023. Arabic regional dialect identification (ardi) using pair of continuous bag-of-words and data augmentation. *International Journal of Advanced Computer Science and Applications*, 14(11).

Reema Khaled AlRowais and Duaa H. Alsaeed. 2023. Arabic stance detection of covid-19 vaccination using transformer-based approaches: A comparison study. *Arab Gulf Journal of Scientific Research*, 42(1):89–107.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resources Association.

Javier Borge-Holthoefer, Walid Magdy, Kareem Darwish, and Ingmar Weber. 2015. Content and network dynamics behind egyptian political polarization on twitter. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW)*, pages 700–711. ACM.

Amey Hengle, Aditya Kshirsagar, Tanmoy Chakrabarty, and Kathleen McKeown. 2021. Combining context-free and contextualized representations for arabic

Table 3: Performance of our system on Arabic data for Subtask 1 (Polarization Detection). All values are reported for the Arabic language only.

Set	Accuracy	Precision	Recall	F1 (Binary)	F1 (Macro)	F1 (Micro)
Development	0.7929	0.7703	0.7600	0.7651	0.7900	0.7929
Blind Test	0.8330	0.8230	0.7988	0.8107	**0.831**	**0.833**

sarcasm detection and sentiment identification. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 207–214, Online. Association for Computational Linguistics.

Lana Kazkaz. 2022. Role of new opinion leaders on social media in political and religious polarization (jordanian case study). *Lebanese Science Journal*, 21(2):233–251.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acarürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Elena Tutubalina, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Tanmoy Chakraborty, and 15 others. 2026a. SemEval-2026 Task 9: Detecting Multilingual, Multicultural and Multievent Online Polarization. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, CA, USA. Association for Computational Linguistics.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acartürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Simona Frenda, Alessandra Teresa Cignarella, Elena Tutubalina, Oleg Rogov, Aung Kyaw Htet, and 24 others. 2026b. Polar: A benchmark for multilingual, multicultural, and multi-event online polarization. *Preprint*, arXiv:2505.20624.

Ingmar Weber, Venkata Rama Kiran Garimella, and Alaa Batayneh. 2013. Secular vs. islamist polarization in egypt on twitter. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 290–297. ACM.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.