EGCSS at StanceNakba Shared Task: Cross-Topic Arabic Stance Detection for Two Middle East Issues

Asmaa Qindeel, Toka Khaled, Batool Balah, Eman Elrefai, Mahmoud Fawzi
Cairo University, Al-Azhar University, Independent Researcher,
Alexandria University, The University of Edinburgh
asmaaqandeel2511@gmail.com, tokakhaled98@gmail.com, batoolnajeh@gmail.com,
eman.lotfy.elrefai@gmail.com, m.f.g.ibrahim@sms.ed.ac.uk

Abstract

Stance detection continues to be an important task sitting at the intersection of Natural Language Processing (NLP) and Computational Social Science (CSS). In this work, we evaluate how different variations of BERT models perform on the cross-topic form of the task. In particular, we inspect their performance on the second subtask of the shared task *StanceNakba 2026*, where two topics are included, namely *Arab Normalization with Israel* and *The Presence of Refugees in Arab Countries*. We find that the best-performing model was **bert-base-arabertv02-twitter**, and we further improve its performance by providing context about the topic during the training phase, achieving an F1-score of **0.86** and ranking second among the participating teams.

Keywords: Stance detection, Arabic, Middle East, political NLP, BERT

1. Introduction

The Middle East is one of the most volatile regions in the world (Fawzi et al., 2026b) where social media is an important space for information exchange (Fawzi, 2022). This situation increases the importance of online arguments about political issues there as they develop causing polarization and fueling conflicts like the recent 2026 Iran War. In this work, we evaluate the performance of different variants of BERT models on the task of cross-topic stance detection that is presented within the *StanceNakba 2026* shared task (Aldous et al., 2026). The two topics introduced are both related to the Middle East where the first topic is the *Arab Normalization with Israel* and the second one is *The Presence of Refugees in Arab Countries*. It is required to develop a system that takes a short social media thread in Arabic about either topic and determines the stance of this thread as being either `Favor`, `Against`, or `Neither`. Our proposed system[1] relies on tuning a BERT (Bidirectional Encoder Representations from Transformers) model that has been trained beforehand on Arabic Twitter data. In addition to hyperparameter tuning and input preprocessing, we also added more context to the input during the training phase to make the model more aware of the topic to which the stance is directed. We achieve a macro F1 score of 0.8607 on the leaderboard and the second rank among the participating teams. We also perform further performance analysis and find that the model performs significantly better on the *Refugees* topic than the *Normalization* topic. In addition, we find that the misclassified samples in the normalization topic are mostly `Against` or `Favor` samples that are misclassified as `Neither` indicating more difficulty in identifying a clear stance towards the topic.

[1]The code can be found here: https://github.com/AsmaaQ25/StanceNakba2026

2. Background

2.1. Task Description

The *StanceNakba 2026* Shared Task focuses on stance detection in politically polarized social media discourse. In this study, we participate in **Task B: Cross-Topic Stance Detection**. The objective is to develop a single model capable of predicting the stance of an Arabic social media post toward a given topic out of a range of multiple topics. Each instance consists of a post–topic pair as input, and the system must output one of three stance labels: `Favor`, `Against`, or `Neither`. The task is framed as a multi-class classification problem and evaluated using macro-averaged F1-score. The dataset of Task B is a subset of the MARASTA dataset (Charfi et al., 2024) and it contains 1,205 manually annotated Arabic social media posts spanning two socio-political topics: *Arab Normalization with Israel* and *Refugee/Immigrant Presence in Jordan*. There are 577 threads about the *Normalization* topic and 628 threads about the *Refugees* topic. The dataset is balanced across stance categories to mitigate bias toward majority classes. The language reflects informal Arabic social media discourse, including dialectal variation and implicit stance expressions, which increase task difficulty.

2.2. Related Work

Stance detection has been widely studied in NLP, particularly in the context of social media. The SemEval-2016 Task 6 benchmark (Mohammad

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 294–298
11 May 2026.

et al., 2016) established stance detection toward specific targets in English tweets. Subsequent work extended the problem to cross-target and multilingual settings, highlighting the challenge of generalization across topics (Vamvas and Sennrich, 2020). Some surveys further categorize stance detection approaches, emphasizing transfer learning and transformer-based architectures (Hardalov et al., 2022).

Recent work has shown growing interest in Arabic stance detection, especially with the introduction of dedicated datasets and shared tasks. StanceEval 2024 (Alturayeif et al., 2024) represents the first large-scale Arabic stance-detection shared task and provides a benchmark dataset together with a comparative evaluation of multiple systems. In addition, the ArabicStanceX dataset (Alkhathlan et al., 2025) was recently introduced as a multi-topic Arabic social-media dataset, where the authors evaluated stance-detection models using transformer-based architectures. Other work, such as (Alkhraiji and Azmi, 2025) focused mainly on traditional Machine Learning models applied to Arabic tweets.

Despite these efforts, most previous work focuses either on dataset construction or on classical machine-learning approaches, while fewer studies explicitly address stance detection in short informal Arabic tweets using transformer-based models. In this work, we focus specifically on short informal tweet text and evaluate numerous BERT-based transformer models in a cross-topic stance-detection setting.

3. System Overview

We chose to work with BERT architecture due to its reported simplicity and efficiency (Farha and Magdy, 2021). The utilization of these pretrained transformers only requires adding a classification head to the base backbones. After filtering, we consider several flavors of the BERT architecture (Farha and Magdy, 2021; Alturayeif et al., 2022).

- bert-base-arabertv02
- MARBERT
- MARBERTv2
- bert-base-arabic-camelbert-da
- bert-base-qarib
- Mawqif
- bert-base-arabertv02-twitter

In the beginning, we tried optimizing generic models for our task by finetuning them using the provided dataset. The main challenge was the size of the dataset as it is not large enough to tweak the models significantly. Hence, we decided for a customized model for the task, which is **bert-base-arabertv02-twitter**. It is already trained on social media tweets, which are similar to our dataset. As a result, it needs less training than generic models. As shown in table 1, this model outperforms other BERT variants.

Model	F1-score
bert-base-arabertv02	0.73
MARBERT	**0.79**
MARBERTv2	0.78
camelbert-da	0.73
bert-base-arabertv02-twitter	**0.79**
QARIB	0.74
Mawqif	0.74

Table 1: Initial test results of different BERT-based models on the validation set using the starter code.

Nevertheless, the reported performance was still not optimal so we started tuning this model to get better results. we have frozen the model encoder layers so that they preserve the following: Core Arabic token understanding, Basic morphology, and syntax representation. This still allows final layers to adapt to our classification task. The starter code provided by the task organizers served as our baseline. It provides the necessary phases we needed including data preprocessing and evaluation metrics. Hence, data preprocessing includes: removing diacritics (tashkeel), tatweel, URLs, and mentions; normalizing alef variants and yaa variants and taa variants as well as spaces.

3.1. Data Augmentation

On our last attempts we tried data augmentation by three methods;

1. Back Translation
2. Generated Samples
3. ArabicStanceX dataset

The first method was executed by translating the Arabic text to english and back using two different LLMs. The second method was done using Claude to generate 300 tweets about the subject. All three methods to increase our data did not achieve any improvements.

4. Experimental Setup

Initially, we started testing several models on the starter code to find a good starting point. The best BERT backbones found were the **MARBERT** model and the **bert-base-arabertv02-twitter**. Afterwards, we started changing the starter code, we

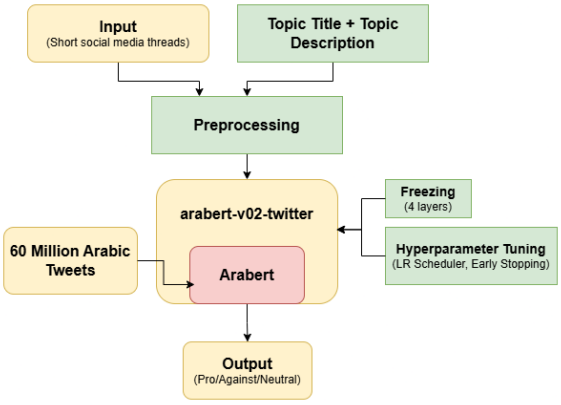

Figure 1: The system diagram. Blocks in green are contributed by us.

experimented with the following techniques: partial fine-tuning by freezing layers, different preprocessing, adding topic description was helpful (Alkhathlan et al., 2025), different hyper-parameters, and extra data using data augmentation which was inefficient. After the labeled validation set was revealed, we used it for cross-validation and early stopping to tune the epochs number. The following hyperparameters yielded the best results:

- **Number of Epochs:** 10
- **LR Scheduler:** linear
- **Learning Rate:** 2e-5
- **Maximum Length:** 64
- **Number of frozen Layers:** 4

دعم بناء علاقات بين الدول العربية ودولة اسرائيل ، دعم إقامة اللاجئين الأجانب داخل الدولة ،

Figure 2: Topic descriptions provided to the model as input during the training phase.

Figure 1 shows the sequence of steps in our system. As shown, **arabertv02-twitter** is a variant of Arabert that was trained with 60 million Arabic tweets. Figure 2 shows the descriptions of the topics that we added to their titles during the learning process to provide the model with more context which turned out to improve the performance.

5. Results

The shortlisted models that we tried to improve further are shown in table 2. The **bert-base-arabertv02** still outperformed other models on the test set; hence, the reported results in the following sections of this paper are based on it. The model achieved an F1-score of `0.8616` and an accuracy of **0.8619** on the test set. This was the second best performing model in the competition only 1% behind the best ranking model.

Model	F1-score
bert-base-arabertv02-twitter	0.86
bert-large-arabertv02-twitter	0.86
Marbert-v2	0.83

Table 2: Final results on test set after parameter tuning.

5.1. Ablation Analysis

Method	Improvement
LR Scheduler	+3%
Freezing Layers	+3%
Max Length	+2%
Early Stopping	+1%
Topic Description	+1%

Table 3: Impact of each method on the fine-tuned model performance.

Table 3 demonstrates the contribution of each tuning step. Finetuning the learning rate scheduler and the early stopping with the labeled validation set resulted in the most positive outcomes. It is important to note that the improvements achieved by these techniques do not necessarily add up together. In other words, the total achieved improvement after incorporating these techniques is less than the sum of the individual contribution.

5.2. Performance Analysis

Class	Precision	Recall	F1-score	Support
Against	0.88	0.89	**0.89**	**65**
Neutral	0.77	0.88	0.82	56
Pro	**0.96**	0.82	0.88	60
Macro Avg	0.87	0.86	0.86	181

Table 4: Classification Report of test set on the final model

To better understand the results of our final model, we investigate further its performance per class and per topic. Table 4 demonstrates that The model performed best on the `Against` samples, while the worst performance was associated with the `Neutral` class. This suggests that the model is able to efficiently detect the presence of negative stances and that it finds neutral samples the most confusing.

Next, we check the performance of the model per topic. As the metrics in tables 6 and 5 illustrate, the

Class	Precision	Recall	F1-score	Support
Against	0.92	0.77	**0.84**	30
Neutral	0.67	**0.90**	0.77	**31**
Pro	**0.95**	0.72	0.82	25
Macro Avg	0.84	0.80	0.81	86

Table 5: Classification Report of the "Normalization" topic (test set)

Class	Precision	Recall	F1-score	Support
Against	0.85	**1.00**	0.92	**35**
Neutral	0.95	0.84	0.89	25
Pro	**0.97**	0.89	**0.93**	**35**
Macro Avg	0.93	0.91	0.91	95

Table 6: Classification Report of the "Refugees" topic (test set)

model performs remarkably better on the topic of *Refugees* than the topic of *Normalization*, although the difference in the number of samples between the two topics was not big. This difference suggests that the complexity of discussions about the topic of *Normalization* can be higher than those related to the topic of *Refugees*.

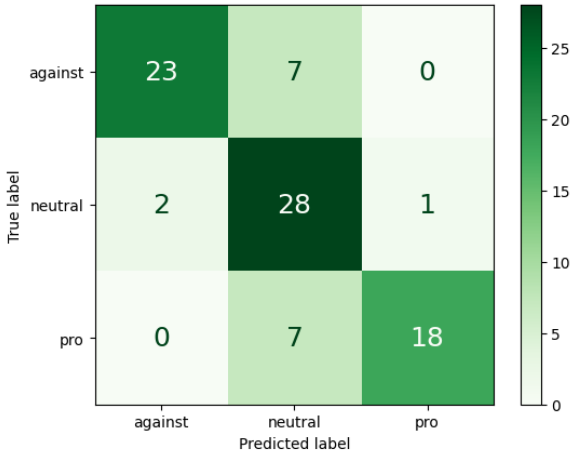

Figure 3: Confusion Matrix of the "Normalization" topic

Finally, we check the confusion matrix for each topic on the test set. We see that the model achieved a 100% recall for the `Against` on the *Refugees* topic. For the *Normalization* topic, we find that that most of the samples where the model was confused belonged to either `Against` or `Pro` classes and were misclassified as `Neutral`.

6. Limitations

This study has several limitations that should be acknowledged. First, the size of the dataset is relatively small (approximately 1,500 samples), which is not sufficient for full fine-tuning of large transformer models. As a result, the model had to be

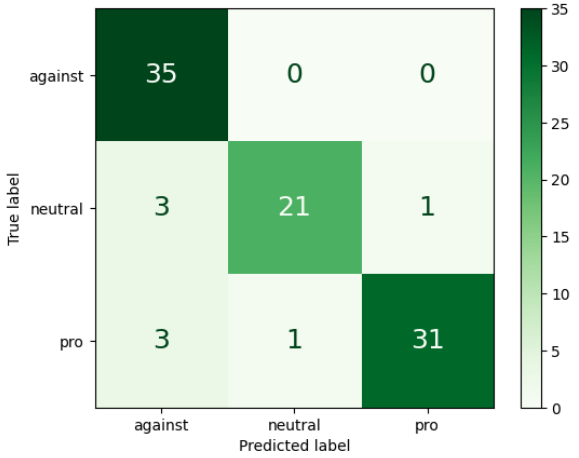

Figure 4: Confusion Matrix of the "Refugees" topic

partially frozen during training, which may have limited its ability to fully adapt to the stance-detection task. Second, explicit stance in Arabic tweets is often difficult to interpret, even for human readers. In many cases, the stance cannot be reliably inferred from the tweet alone and may depend on the author's previous opinions, background context, or implicit references that we do not have. This makes the task inherently challenging and increases the likelihood of ambiguous or misleading samples in the dataset. This challenge can be addressed by using network features for the processing of Arabic social data as earlier works did (Fawzi and Magdy, 2024; Fawzi et al., 2026a).

7. Conclusion

In this paper, we presented our system for Task B of the *StanceNakba 2026* shared task on cross-topic Arabic stance detection. We evaluated several BERT-based models and found that **bert-base-arabertv02-twitter** was best suited for this social media task. We further improved its performance through layer freezing, hyperparameter tuning, and the addition of topic descriptions as contextual input. Our final system achieved an F1-score of **0.86**, ranking second among the participating teams. Future work may explore larger Arabic language models and more effective data augmentation strategies to address the limited dataset size.

8. Bibliographical References

Kholoud Khalil Aldous, Md Rafiul Biswas, Mabrouka Bessghaier, Shimaa Ibrahim, Kais Attia, and Wajdi Zaghouani. 2026. StanceNakba shared task: Actor and topic-aware stance

detection in public discourse. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

Ali Alkhathlan, Faris Alahmadi, Faris Kateb, and Hend Al-Khalifa. 2025. Constructing and evaluating arabicstancex: a social media dataset for arabic stance detection. *Frontiers in Artificial Intelligence*, 8:1615800.

Arwa K. Alkhraiji and Aqil M. Azmi. 2025. Stance detection in arabic tweets: A machine learning framework for identifying extremist discourse. *Mathematics*, 13(18).

Nora Alturayeif, Hamzah Luqman, Zaid Alyafeai, and Asma Yamani. 2024. $stanceeval2024 : The first arabic stance detection shared task$. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 774–782, Bangkok, Thailand. Association for Computational Linguistics.

Nora Saleh Alturayeif, Hamzah Abdullah Luqman, and Moataz Aly Kamaleldin Ahmed. 2022. Mawqif: A multi-label arabic dataset for target-specific stance detection. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 174–184.

Anis Charfi, Mabrouka Ben-Sghaier, Andria Samy Raouf Atalla, Raghda Akasheh, Sara Al-Emadi, and Wajdi Zaghouani. 2024. Marasta: A multi-dialectal arabic cross-domain stance corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11060–11069.

Ibrahim Abu Farha and Walid Magdy. 2021. Benchmarking transformer-based language models for arabic sentiment and sarcasm detection. In *Proceedings of the sixth Arabic natural language processing workshop*, pages 21–31.

Mahmoud Fawzi. 2022. Characterisation of users' behaviours towards fake news through the analysis of their networks.

Mahmoud Fawzi and Walid Magdy. 2024. "pinocchio had a nose, you have a network!": On characterizing fake news spreaders on arabic social media. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–20.

Mahmoud Fawzi, Björn Ross, and Walid Magdy. 2026a. Fabricating holiness: Characterizing religious misinformation circulators on arabic social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 20.

Mahmoud Fawzi, Björn Ross, and Walid Magdy. 2026b. "god says we are right!": The interplay between religion and propaganda on arabic social media. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis- and disinformation identification. *ACM Computing Surveys*, 55(11).

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.

Jannis Vamvas and Rico Sennrich. 2020. X-stance: A multilingual multi-target dataset for stance detection. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 151–163.