Yafa at StanceNakba: Actor-Level Stance Detection Using Cross-Lingual Approach

Tasnim Zayet, Osama Hamed*, Tasneem Duridi
Department of Computer Science, Birzeit University
Computer Systems Engineering Department, Palestine Technical University - Kadoorie
Department of Computer Science, Palestine Technical University - Kadoorie
{Birzeit, Tulkarm}, West Bank - Palestine
tzayet@birzeit.edu, {osama.hamed, tasneem.duridi}@ptuk.edu.ps
*Corresponding author.

Abstract

This paper addresses the problem of actor-level stance detection in English social media posts concerning the Palestinian issue, a subtask of the StanceNakba-2026 Shared Task. The objective is to classify posts into one of three categories: Pro-Palestine, Pro-Israel, or Neutral, which is more challenging than the traditional favor/against/neutral formulations. This study uses a dataset comprising 1,401 posts, collected from X (formerly Twitter) after October 7, 2023, and annotated with one of the three stance labels. As Yafa's Team, we tried to solve this problem using BERT-based models, which have proven their superiority in similar tasks. Several BERT-based models were fine-tuned and compared, including ARBERT, MARBERT, and PoliBERTweet, among others. Our winning model is the "MARBERT-Y", where the "Y" comes from Yafa, a MARBERT-based model that has achieved a macro-F1 score of 95% on the test set. We argue this to two main factors: the structured and harsh preprocessing steps applied and the fine-tuning process employed. This indicates that domain-adapted transformer models, i.e., those pretrained on large-scale Twitter data are highly effective for politically stance detection tasks.

Keywords: Nakba, stance detection, multi-class classification, BERT-based models, MARBERT, hyperparameters.

1. Introduction

Social media platforms have grown tremendously in recent years. As a result, military conflicts have also moved into the digital space, shaping individuals' opinions around the world in real time (Duridi et al., 2025b). Additionally, people can access the perspectives of all parties involved in a conflict along with supporting evidence. This has led people to take sides. Over the past two years, the war against Gaza has attracted the world's attention. As a consequence, a vast amount of content has spread all over the internet representing various stances. Some support Israel, others stand with the Palestinians, and some maintain a neutral stance (Duridi et al., 2025a). Specifying and analyzing people's stance in an automatic manner is a challenging and important task. It may mobilize and influence public opinions, help in detecting fake news and analyze trends, especially in elections.

This paper outlines the Yafa's Team (yafateam) submission to Subtask A of the StanceNakba-2026 Shared Task (Aldous et al., 2026) at the NakbaNLP 2026 International Workshop (LREC 2026 Conference). Where we tackle the actor-level English stance detection towards the Palestinian-Israeli conflict. The task is to build a multi-class text classification model that can identify an author's political opinion. In other words, the task is to categorize social media post as Pro-Palestine, Pro-Israel, or Neutral. This task is more challenging than other stance detection tasks where the categories are Favor, Against, or Neutral. For example, see the work by Melhem et al. (2024). To approach this task, we employ several BERT-based classification approaches, which investigate the pre-trained transformer model's ability to capture the deep contextual representations of text. We fine-tune BERT-based models using a task-specific dataset consisting of 1,401 English posts. As requested, we had a data-split with 70/15/15 for training, development and testing respectively. Our model learns to distinguish subtle linguistic cues and stance expressions characteristic of this politically sensitive domain, enabling robust classification of author-level political positions.

This paper offers the following three contributions: (i) we present a single-task learning (STL) BERT-based model that utilizes MARBERT to detect actor-level stance on the **Pro-Palestine**, **Pro-Israel**, or **Neutral** opinions, (ii) we describe the list of hyperparameter (HP) that made MARBERT outperform other used BERT-based models, thus ranking Yafa Team 2nd, and (iii) we provide a detailed analysis and discussion of our best-performing model accross thees three opinions.

The paper is organized as follows: Section 2 presents prior and recent research on stance de-

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 177–181

tection. Section 3 provides a comprehensive analysis of the social media dataset used. Section 4 describes the proposed systems and experimental setup. Section 5 presents and discusses our experimental results. Finally, Section 6 concludes the paper and suggests ideas for future search.

2. Related Work

Actor-level stance detection focuses on analyzing an individual's opinions towards a specific target and it is an STL that has been widely utilized in elections and conflicts (Mohammad et al., 2016).

Previous studies on actor-level stance (Mohammad et al., 2016) addressed stance detection for specific political targets using English tweets with supervised traditional classifiers, specifically Support Vector Machines (SVMs) and convolutional neural networks (CNN) (Vijayaraghavan et al., 2016). An improvement upon these was done using Bidirectional Long Short-Term Memory (Bi-LSTM) that jointly encodes tweet-target representation (Augenstein et al., 2016). Further advancements were proposed for multi-target stance detection using an ensemble neural approach (Sobhani et al., 2017; Siddiqua et al., 2019).

Recent studies using transformer-based models such as BERT improved actor-level stance detection. Conforti et al. (2020) employed BERT which outperformed earlier architectures in both in-target and cross-target setups, while Kawintiranon and Singh (2021) incorporated political knowledge into BERT's masked language model. Additional work explores zero-shot transfer across targets and topics using BERT-style encoders (Allaway and McKeown, 2020).

Domain-adapted models such as PoliBERTweet (Kawintiranon and Singh, 2022), which was pre-trained on 83 million US election tweets, achieved 3-10% F1 gains over general-purpose models. AraBERT and MARBERT (Abdul-Mageed et al., 2021), pre-trained on large-scale Arabic corpora, outperformed multilingual models like mBERT and XLM-RoBERTa across Arabic NLP benchmarks.

Despite this progress, existing work targets western political contexts with classes favor,against and neutral, leaving cross-cultural conflict scenarios underexplored. Moreover, it remains unclear whether linguistic adaptation outweighs domain adaptation for English tweets related to Arabic politics and cultural context. Our work addresses these gaps by benchmarking multiple BERT-based models, showing that Arabic-based models pre-trained on tweets outperformed both multilingual models and domain-adapted English models, establishing that linguistic and cultural alignment are more critical than topical domain relevance for this type of cross-cultural stance detection.

3. Data and Stats

This section presents an overview of the StanceNakba-2026 dataset and provides descriptive statistics to better understand its structure and distribution. The dataset was collected and annotated by the shared task organizers.

3.1. Dataset Overview

The StanceNakba-2026 dataset is a balanced dataset consisting of 1,401 posts, primarily collected from X (formerly Twitter) social media platform, during the period following October 7, 2023. Each post is annotated with one of three stance labels reflecting the author's position toward the ongoing conflict: **Pro-Israel**, **Pro-Palestine**, or **Neutral**. The task is formulated as a multi-class supervised stance detection problem.

The dataset was released in three separate phases to support the ML/AI model development and evaluation, in line with the shared task design. Table 1 summarizes the composition of the dataset across these phases.

Phase	Total	Pro-Pal.	Pro-Isr.	Neut.
Training	980	326	327	327
Validation	210	55	88	67
Test	211	–	–	–

Table 1: Dataset composition and label distribution across splits.

The training set exhibits a balanced distribution across the three stance categories, which supports effective supervised learning and reduces the risk of strong class bias. The label distribution is illustrated in Figure 1.

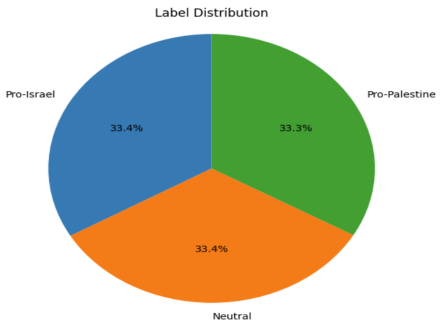

Figure 1: Distribution of stance labels in the training set.

To provide qualitative insight into the nature of the data, Table 2 presents representative examples from each stance category. These examples highlight the diversity in tone, content, and linguistic expression found in the dataset, reflecting the complexity of stance detection in real-world social media discourse.

Label/Pro-	Example
Palestine	Against the genocide that the Zionists commit.
Israel	Finish Hamas. I wish we had a president who would assist Israel to do what needs to be done. Rid the world of these terrorists who murder and violate men, women and children. Enough!
Neutral	The Lord Himself gave that land to the Jews and I will back them till I die.

Table 2: Representative examples from each stance category.

3.2. Data Preprocessing

A systematic preprocessing pipeline was applied to all textual data to improve data quality and maintain consistency throughout the corpus as follows:

- Lowercase normalization: All alphabetic characters were normalized to lowercase to reduce lexical variation and minimize feature space complexity.
- Links normalization: All links were replaced by empty strings.
- User mention Normalization: All user mentions were replaced with uniform text @user.
- Character repetition normalization: character repetitions were removed.
- White space normalization: repeated white spaces were removed; additionally, the tweets were stripped.
- Connecting words normalization: connected words were split using the wordninja library (Anderson, 2019).
- Hashtag normalization: The hashtag notation # was removed, while the hashtag text was kept.

However, punctuation marks and emojis were kept due to their importance in stance detection, as some (e.g., ?,!) can indicate stance. These preprocessing steps produced clean, semantically relevant text optimized for feature extraction and subsequent classification analysis.

4. System Description

This section briefly describes the system we built to take part in the NakbaStance detection Shared Task. Our system is a multi-class classification system that evaluates multiple BERT-based pretrained language models.

4.1. Used Models

4.1.1. XLM-Roberta-base

This is a cross-lingual pretrained language model proposed by (Conneau et al., 2019). It was trained on large-scale multilingual corpora comprising 2.5TB of web-crawled text from 100 languages. The base configuration consists of approximately 125 million parameters, offering a practical trade-off between model capacity and computational cost.

4.1.2. ARBERT and MARBERT

Introduced by (Abdul-Mageed et al., 2021), ARBERT and MARBERT are two Arabic-specific pretrained transformer models that extend the BERT-base architecture, each comprising 12 attention layers, 768 hidden dimensions, and approximately 163 million parameters. The key distinction lies in their pretraining corpora: ARBERT is trained on a 61GB Arabic corpus drawn from Wikipedia, news, and literary sources, whereas MARBERT is pretrained on a large-scale collection of 1 billion tweets (128GB), covering both Modern Standard Arabic and a wide range of dialectal varieties. To accommodate the brevity of tweet content, MARBERT omits the Next Sentence Prediction (NSP) objective during pretraining. Both models surpassed strong multilingual baselines including mBERT and XLM-RoBERTa across different Arabic NLP tasks.

4.1.3. PoliBERTweet

PoliBERTweet extends BERTweet through continued pre-training on 83 million US election tweets (Kawintiranon and Singh, 2022). This specialization improves its handling of political discourse, outperforming BERTweet and other baselines on stance detection and sentiment analysis with 3-10% F1 score gains.

4.1.4. BERT-base-uncased

BERT-base-uncased (Devlin et al., 2019) is the foundational model with 12 layers, 768 hidden units, and 110M parameters. Pre-trained on BookCorpus and Wikipedia using MLM (15% masking) and NSP, it employs WordPiece tokenization (30K vocabulary) and case-insensitive processing.

4.2. Hyperparameter Settings

We fine-tuned UBC-NLP/MARBERT for three-class stance classification using the Hugging Face Trainer framework. The dataset was divided into training (70%), development (15%), and test (15%) splits using stratified sampling to preserve label distribution. A fixed random seed was used to ensure reproducibility.

Optimization was performed using AdamW. The model was trained for a maximum of 8 epochs, with early stopping applied if the macro-averaged F1 score on the development set did not improve for two consecutive epochs. Table 3 describes the list of hyperparameters (HP) applied.

Note that the input sequences were tokenized using the MARBERT tokenizer, i.e., truncated to a maximum length of 128 tokens, and padded to a fixed length. It should be also noticed that no learning rate warm-up or gradient accumulation was applied. Additionally, a customized cross-entropy function with label smoothing was employed to suit the task of multi-class classification and decrease the effect of the confidence problem. Besides label smoothing, the dropout rate in attention layers was increased to 0.2.

Hyperparameter	Value
Learning rate	2×10^{-5}
Weight decay	0.1
Batch size (train / eval)	4 / 4
Maximum sequence length	128
Number of epochs	8
Early stopping patience	2
Optimizer	AdamW

Table 3: Main hyperparameters used for fine-tuning.

4.3. Model Evaluation

Following the organizers' evaluation framework, we use F1-macro metric as the primary one. F1-macro computes the F1 score (i.e. the harmonic mean of precision and recall) for each class independently and averages them equally. This makes it robust against class imbalance.

5. Results & Discussion

As managed by the organizers, the dataset was split into training, development/validation, and testing subsets, allocating 70%, 15%, and 15% of the data, respectively.

During the development phase, our model, i.e., MARBERT-Y was regularly trained on the training split and tested using the development split. It has achieved an F1-macro of 88%, even without applying a harsh pre-processing or changing a lot of HP. However, the validation gave indications that this model is promising.

During the testing phase, our model achieved an F1-macro of 95%. We argue this to the applied harsh pre-processing described in Subsection 3.2, and the adopted set of configurable HP, see subsection 4.2. Table 4 compares the performance of the used BERT-based models according to F1-macro, and showing that MARBERT stood first. Consequently, we call it as MARBERT-Y, where the "Y" stands for Yafa.

Model	Acc.	Prec.	Rec.	F1
ArBert	0.94	0.94	0.94	0.94
MARBERT	0.95	0.95	0.95	**0.95**
PoliBERTweet	0.91	0.91	0.91	0.90
BERT-base-uncased	0.91	0.91	0.91	0.91
XLM-Roberta-base	0.91	0.92	0.91	0.91

Table 4: Performance comparison of BERT-based models.

Initially, we hypothesized that domain-specific political models would achieve superior performance. However, PoliBERTweet underperformed compared to Arabic-specialized models. This indicates that linguistic compatibility is more critical than domain relevance. A translation-based approach to Arabic also proved ineffective due to the semantic distortions introduced during automatic translation. Since the dataset discusses Arabic political topics in English but contains terms whose meaning remains tightly bound to their original linguistic and cultural context, even after translation, we adopted Arabic-specialized models: ArBERT and MARBERT. Although, multilingual models such as XLM-RoBERTa-base achieved competitive results, they still fell short. Finally, ArBERT and MARBERT achieved the best performance with scores of 0.94 and 0.95 respectively. This demonstrates the advantage of large-scale Arabic pre-training, particularly for dialectal expressions, as discussed previously in Section 4.

6. Conclusion

In this paper, the Yafa's team (ranked 2nd on the leaderboard) tried various BERT-based models for the actor-level stance detection, subtask (A) of the NakbaStance-2026 Shared Task. Our best-performing model is the MARBERT-Y. It gets it's name from the MARBERT pretrained model, and the "Y" from Yafa. It has achieved a macro-F1 of 95%. We attribute this model's superiority to three factors: the large-scale collection of tweets used during pretraining, the rigorous preprocessing steps applied, and the set of tuned hyperparameters.

7. Acknowledgements

The authors would like to thank their institutions for their support: Birzeit University and Palestine Technical University - Kadoorie.

8. Bibliographical References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, et al. 2021. Arbert & marbert: Deep bidirectional transformers for arabic. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing, volume 1*, pages 7088–7105.

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.

Emily Allaway and Kathleen McKeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. *arXiv preprint arXiv:2010.03640*.

Derek Anderson. 2019. wordninja 2.0.0.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. *arXiv preprint arXiv:1606.05464*.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won't-they: A very large dataset for stance detection on twitter. *arXiv preprint arXiv:2005.00388*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1*, pages 4171–4186.

Tasneem Duridi, Lour Atwe, Areej Jaber, Eman Daraghmi, and Paloma Martínez. 2025a. Detection of propaganda and bias in social media: A case study of the israel-gaza war (2023). In *2025 International Conference on New Trends in Computing Sciences (ICTCS)*, pages 204–210. IEEE.

Tasneem Duridi, Areej Jaber, and Paloma Martínez. 2025b. Arabic hate speech detection based on bert models variants. *Egyptian Informatics Journal*, 32:100845.

Kornraphop Kawintiranon and Lisa Singh. 2021. Knowledge enhanced masked language model for stance detection. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 4725–4735.

Kornraphop Kawintiranon and Lisa Singh. 2022. Polibertweet: a pre-trained language model for analyzing political content on twitter. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7360–7367.

Anas Melhem, Osama Hamed, and Thaer Sammar. 2024. Tao at stanceeval2024 shared task: Arabic stance detection using arabert. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 842–846.

Saif 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.

Umme Aymun Siddiqua, Abu Nowshed Chy, and Masaki Aono. 2019. Tweet stance detection using an attention based neural ensemble model. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1*, pages 1868–1873.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557.

Prashanth Vijayaraghavan, Ivan Sysoev, Soroush Vosoughi, and Deb Roy. 2016. Deepstance at semeval-2016 task 6: Detecting stance in tweets using character and word-level cnns. *arXiv preprint arXiv:1606.05694*.