shroukgbr at StanceNakba Shared Task: Transformer-Based Ensemble Learning for Actor-Level Stance Detection in Palestinian–Israeli Social Media Discourse

Shrouk Anwar Gabr, Mohamed Ibrahim Ragab

School of Information Technology and Computer Science (CIS)
Nile University, Giza, Egypt
SGabr@nu.edu.eg, MoRagab@nu.edu.eg

Abstract

Stance detection has become an essential task for understanding political discourse on social media, particularly in highly polarized contexts where sentiment alone is insufficient to capture author intent. This study addresses stance classification in discussions related to the Palestinian–Israeli conflict by developing transformer-based and ensemble learning approaches for three-class classification: Pro-Palestine, Pro-Israel, and Neutral. Using the StanceNakba 2026 Shared Task dataset, we fine-tune multiple pretrained transformer models, including MARBERT, ARBERT, BERT, RoBERTa, and DeBERTa, and evaluate their performance using stratified cross-validation with macro F1-score as the primary metric. In addition to individual model evaluation, a weighted ensemble combining BERT, RoBERTa, and DeBERTa is proposed to leverage complementary contextual representations. Experimental results show that the ensemble model achieves the best performance with an accuracy and macro F1-score of 0.8905, outperforming specialized Arabic models while maintaining strong class-wise balance.

Keywords: Actor-level stance detection, Weighted cross-entropy, ARABERT, MARBERT, Transformer-based text classification, Political stance modeling, Cross-validation ensembling.

1. Introduction

The advent of social media has forever changed the manner in which people communicate regarding political issues. Online political discussions can be highly emotionally charged and express strong opinions about certain types of people, including political actors, organizations, and political movements. In order to have a complete understanding of these types of discussions, simply performing a general sentiment analysis is not enough; we also need to determine actor-level stance detection, which will provide an indication of whether a reader supports or opposes an actor or if they remain neutral toward that actor in the text being analyzed. Stance detection refers to the determination of a writer's position on a specific target (Gera and Neal, 2025). While sentiment analysis simply looks at the polarity of the sentiments expressed in the text (i.e., whether the sentiment expressed is positive or negative), stance detection also includes which actor or target the sentiment is being expressed towards (Garg and Caragea, 2024).

Understanding public opinion, tracking misinformation, and analyzing political debates benefit from stance detection (Garg and Caragea, 2024), (Ahmad and Kakudi, 2025). Most research pertains to topics, while previous research examined how tweets discuss the Palestinian-Israeli conflict using three categories: pro-Palestine, pro-Israel, or neutral (Imtiaz et al., 2022). We take a different approach to stance detection by creating one unified model that will assign stance classification (pro-Palestine, pro-Israel, or neutral) based on an author's complete political stance on the Palestinian-Israeli conflict, evaluated using the StanceNakba 2026 Shared Task dataset and benchmark (Aldous et al., 2026). This approach secured first place on the Codabench leaderboard in both the development and final evaluation phases, demonstrating its robustness and effectiveness in real-world stance detection scenarios. To support reproducibility, the implementation is publicly available at: `https://github.com/shroukgbr89/StanceNakba-2026-Stance-Detection`, including executable Google Colab notebooks.

The rest of the paper is organized as follows: Section 2 shows related work on the propaganda classification problem. Section 3.1 describes the dataset and preprocessing techniques used in this study, and Section 3.2 details the transformer models and ensemble frameworks implemented. Section 4 presents the experimental results, including individual model performance and ensemble outcomes. Finally, Section 5 discusses the findings, limitations, and potential directions for future research.

2. Related Work

This section surveys recent developments in closely related areas, including propaganda detection, political bias detection, extremism detection, and multilingual misinformation detection. These works establish the contextual and methodological foundation for multilingual classification research, out-

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 280–284

lining key challenges, commonly used approaches, and existing limitations that inform and motivate the present study's contributions.

The study (Ragab et al., 2025) uses the SinaLab FigNews 2024 multilingual dataset, consisting of 13,500 Facebook posts across five languages: Arabic, English, Hebrew, French, and Hindi (2,400 posts per language). The dataset includes original text, machine translations, and bias-related annotations, with the main classification target being a four-class. Due to strong class imbalance, the authors applied oversampling to create a fully balanced dataset with 7,098 instances per class. Three transformer models—mBERT, XLM-RoBERTa, and mT5—were evaluated on both imbalanced and balanced data. On imbalanced data, mT5 significantly outperformed others (98.86% accuracy, F1=0.9882), while mBERT (53%) and XLM-RoBERTa (69.7%) struggled. After balancing, all models improved substantially: mT5 achieved 99.61% accuracy and F1=0.9962 with the lowest training and validation loss, demonstrating superior generalization; Overall, results confirm that mT5 is the most robust and effective.

The study (Ahmad et al., 2020) investigates ensemble machine learning methods for automatic fake news detection using a publicly available dataset of news articles labeled as fake or real. The textual data undergo standard NLP preprocessing steps, including tokenization, stopword removal, and TF-IDF feature extraction, and is split into training and testing sets with a relatively balanced class distribution to ensure fair evaluation. Results show that ensemble models, such as random forest and voting-based approaches, consistently outperform individual classifiers like decision tree, Naïve Bayes, SVM, and KNN across accuracy, precision, recall, and F1-score metrics. The ensemble methods achieve the highest performance, exceeding 90% accuracy, demonstrating stronger robustness and generalization compared to single-model classifiers. The study (Fadri et al., 2026) analyzes public sentiment on the Gaza conflict using a dataset of 2,175 tweets collected from the X platform and processed through standard NLP preprocessing steps, including cleaning, tokenization, stemming, and TF-IDF feature extraction. Tweets were labeled into positive, negative, and neutral classes, creating a high-dimensional textual dataset for sentiment classification. Three models—MLP, XGBoost, and logistic regression—were evaluated using different train–test splits. Logistic regression achieved the best performance with 73.17% accuracy, followed by XGBoost (70.18%), while MLP showed the lowest results (69.27%). The results demonstrate that simpler linear models are more effective than complex models for small, sparse social media datasets represented with TF-IDF features.

Gap Analysis: Despite progress in misinformation detection and multilingual text classification, key gaps remain. Many studies focus on improving model performance but rely on balanced, controlled datasets that do not reflect real-world, imbalanced social media data. Additionally, research often treats tasks like sentiment analysis and fake news detection separately, failing to capture the complexity of multilingual propaganda and bias, which require deeper cross-lingual and cultural understanding. While traditional machine learning methods lack the ability to model semantic relationships, transformer-based approaches are typically evaluated on performance rather than robustness and generalization.

3. Materials and Methods

3.1. The used dataset

This study's datasets were taken from a public discourse dataset regarding the Palestinian–Israeli Conflict that the StanceNakba 2026 Shared Task provided for identifying stances. We participated in Subtask A: Actor-Level Stance Detection. The dataset consists of 1,401 labeled (Target) samples as part of the main training data file; there were additional development and test splits from the organizers of the task. The authors created a corpus of data using a 70/15/15 split to allow for training and evaluating the model.

As shown in 1, the dataset utilized in this study was partitioned into three distinct sets for model development and evaluation. The training set contains 980 labeled instances used for model learning. An additional 210 labeled samples served as the validation set for performance monitoring and hyperparameter tuning. Finally, the test set provided consists of 211 unlabeled instances for which final predictions were generated. There are three main columns on each data file: id, text, and stance-related label columns. An ID column that is a unique identifier for each text. The text column contains the original English language related to how people feel about different aspects of the conflict. For the training and validation sets, a categorical label—Pro-Palestine, Pro-Israel, or Neutral—is also provided, defining the stance classification task.

A particular focus of our analysis is the distribution of stance labels within the training dataset. As can be seen in Figure 1, the three classes are almost equally distributed across the dataset: there are 327 instances of Pro-Israel, 327 instances of Neutral, and 326 instances of Pro-Palestine, for a total of 980 samples. Balanced distributions mitigate any class bias in the training process, allowing for equitable assessment through all classification

Table 1: Sample instances from the StanceNakba 2026 dataset (Subtask A)

Stance Label	Text
Pro-Palestine	may god come in revenge plestine i will not be staying silent while a genocide is unfolding in the lan dof palestine.
Pro-Israel	I pray that our God will give her peace. Such a terrible thing that she had to endure. God please continue to protect Israel and it peoples!
Neutral	It is horrible what the children suffer because of Islamic aggression.

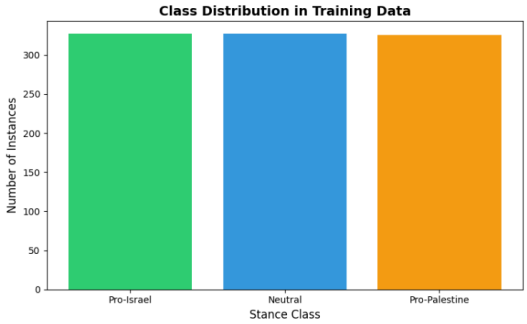

Figure 1: Distribution of instances in the stance column across the dataset.

categories.

3.2. Methodology

In this study, we implemented state-of-the-art multilingual models to classify propaganda and bias in the war between Gaza and Israel. The models include MARBERT, ARABERT, and BERT models.

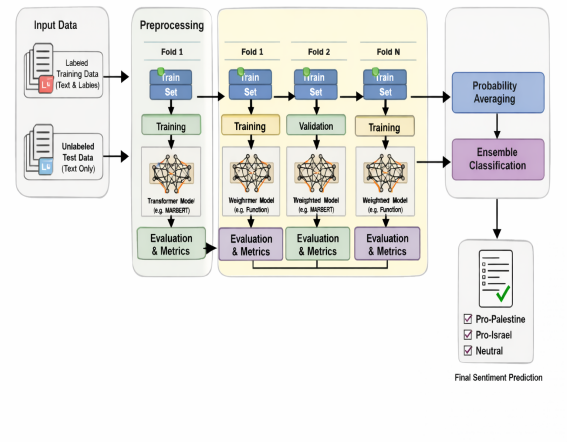

Figure 2: Propaganda Detection Architecture Utilizing MARBERT, ARABERT, and BERT models.

3.2.1. Data Preprocessing

As shown in 2 The raw textual data are first subjected to a series of preprocessing operations to improve data quality and ensure consistency. These steps include removing duplicates and irrelevant entries, text normalization, lowercasing, punctuation removal, and elimination of URLs and special characters. Stopwords are removed to reduce noise, while tokenization is applied to split text into meaningful linguistic units. Stemming, or lemmatization, is then performed to reduce words to their base forms, helping minimize vocabulary sparsity and improving generalization. After preprocessing, textual data are transformed into numerical representations suitable for machine learning algorithms. A term-weighting approach based on Term Frequency–Inverse Document Frequency (TF-IDF) is employed to capture the importance of words within documents relative to the entire corpus. This process produces a high-dimensional, sparse feature matrix that encodes semantic relevance while reducing the influence of commonly occurring but less informative terms.

3.2.2. Model Development Pipeline

The extracted features are used to train multiple supervised machine learning models. The architecture allows experimentation with different classifiers to analyze performance variations across learning paradigms. Each model learns patterns from the training data to distinguish between target classes based on textual characteristics. Hyperparameters are configured to ensure stable learning behavior and to prevent overfitting. The dataset is divided into training and testing subsets using multiple train–test split ratios to evaluate model robustness under different data availability scenarios. During training, models learn decision boundaries from the feature vectors, while validation is performed on unseen data to assess generalization capability.

4. Results and Discussion

4.1. Experiment Setup

We used transformer-based models to determine each actor's stance on the Palestinian–Israeli conflict. The experiment involved data preprocessing, model fine-tuning, and evaluation. All experiments were conducted in a cloud environment using NVIDIA Tesla P100 and NVIDIA T4 GPUs via Google Colab.[1] These accelerators enable efficient fine-tuning of large transformer models such as DeBERTa-v3 and MARBERT. The implementation

[1] https://colab.research.google.com/

was carried out using Hugging Face Transformers [2] and PyTorch [3], along with supporting libraries including Scikit-learn, Pandas, and NumPy. Models were trained using the AdamW optimizer with a linear learning rate scheduler and warmup. Hyperparameters varied across models but generally included a learning rate in the range of 1.5×10^{-5} to 2×10^{-5}, batch sizes of 8–16, and training for 6–10 epochs. The maximum sequence length ranged between 256 and 320 tokens depending on the model.

4.2. Models Evaluation

The preprocessing phase ensured data quality and model compatibility through several steps. Texts were cleaned by removing extra whitespaces and handling missing values. Stance labels were encoded into numerical form for classification. Tokenization was performed using Hugging Face's AutoTokenizer with padding and truncation, and sequence lengths ranged from 192 to 320 tokens. Finally, the data was converted into PyTorch tensors and split into training, validation, and test sets.

Our proposed system achieved **1st place** on the official Codabench leaderboard in both the development and final evaluation phases of the StanceNakba 2026 Shared Task, ranking first among all participating teams. In this section, we evaluate three general-purpose pretrained models, BERT, RoBERTa, and DeBERTa, as well as two specialized models, MARBERT and ARBERT, using stratified cross-validation and macro F1-score as the primary evaluation metric.

Table 2 summarizes the performance of the evaluated models using precision, recall, and F1-score, with macro F1-score serving as the primary evaluation metric. The ensemble model achieved the best overall performance, reaching an accuracy and macro F1-score of 0.8905, outperforming MARBERT (0.88) and ARBERT (0.86). This improvement highlights the benefit of combining multiple transformer architectures to enhance generalization and prediction robustness. Class-wise results show that the Pro-Israel class achieved the highest performance across all models, indicating that strongly polarized content is easier to detect due to clearer linguistic cues. The pro-Palestine class also demonstrated stable performance, with consistently high recall values across models. The Neutral class proved most challenging, as lower recall suggests models frequently confuse neutral statements with partisan content. Overall, the results indicate that ensemble learning improves stance classification performance by leveraging complementary model representations. However, accurately identifying neutral discourse remains a key challenge, highlighting an important direction for future research.

Table 2: Performance comparison of evaluated models using Precision (P), Recall (R), and F1-score.

Model	Class	P	R	F1	Sup
Ensemble	Neutral	0.9206	0.8286	0.8722	70
	Pro-Israel	0.9286	0.9286	0.9286	70
	Pro-Palestine	0.8312	0.9143	0.8707	70
	Acc.		0.8905		
	Macro Avg	0.8935	0.8905	0.8905	210
MARBERT	Neutral	0.92	0.78	0.84	327
	Pro-Israel	0.87	0.94	0.90	327
	Pro-Palestine	0.86	0.91	0.88	326
	Acc.		0.88		
	Macro Avg	0.88	0.88	0.88	980
ARBERT	Neutral	0.90	0.76	0.82	327
	Pro-Israel	0.85	0.94	0.89	327
	Pro-Palestine	0.84	0.89	0.87	326
	Acc.		0.86		
	Macro Avg	0.87	0.86	0.86	980

5. Conclusion

This study investigated transformer-based and ensemble learning approaches for stance detection in political discourse related to the Palestinian–Israeli conflict. Multiple pretrained models, including MARBERT, ARBERT, BERT, RoBERTa, and DeBERTa, were evaluated using standardized preprocessing, fine-tuning strategies, and macro F1-score as the primary evaluation metric. Experimental results demonstrated that the proposed ensemble model achieved the highest overall performance, surpassing individual specialized models by effectively combining diverse contextual representations and reducing model-specific biases. The analysis further showed that strongly polarized stances are more easily identified than neutral content, as neutral statements often contain implicit or ambiguous linguistic cues that resemble partisan expressions.

5.1. Limitations and Future Work

Despite the promising results, this study has several limitations. First, the dataset size and class distribution, particularly for neutral statements. Second, the analysis was constrained to textual content without considering multimodal signals such as images, videos, or network interactions. Finally, computational resources restricted extensive experimentation with larger transformer architectures or fully multilingual models, potentially affecting performance on cross-lingual content.

Future research could address these limitations through several avenues. Increasing dataset size and diversity, particularly for neutral and ambiguous statements, would improve model generalization.

[2] https://huggingface.co/transformers/
[3] https://pytorch.org/

6. References

Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf, and Muhammad Ovais Ahmad. 2020. Fake news detection using machine learning ensemble methods. *Complexity*, 2020(1):8885861.

Mahmoud Ahmad and Habeebah Kakudi. 2025. Stance detection on nigerian 2023 election tweets using bert: A low-resource transformer-based approach. In *Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025)*, pages 54–63.

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.

Agil Irman Fadri, Nur Futri Ayu Jelita, Diamond Dimas Bagaskara, and Raudiatul Zahra. 2026. Sentiment analysis of public opinion on the gaza conflict using machine learning. *Public Research Journal of Engineering, Data Technology and Computer Science*, 3(2):138–148.

Krishna Garg and Cornelia Caragea. 2024. Stanceformer: Target-aware transformer for stance detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4969–4984.

Parush Gera and Tempestt Neal. 2025. Deep learning in stance detection: A survey. *ACM Computing Surveys*, 58(1):1–37.

Arsal Imtiaz, Danish Khan, Hanjia Lyu, and Jiebo Luo. 2022. Taking sides: Public opinion over the israel-palestine conflict in 2021. *arXiv preprint arXiv:2201.05961*.

Guan-Tong Liu, Yi-Jia Zhang, Chun-Ling Wang, Ming-Yu Lu, and Huan-Ling Tang. 2024. Comparative learning based stance agreement detection framework for multi-target stance detection. *Engineering Applications of Artificial Intelligence*, 133:108515.

Mohamed Ibrahim Ragab, Ensaf Hussein Mohamed, and Walaa Medhat. 2025. Multilingual propaganda detection: Exploring transformer-based models mbert, xlm-roberta, and mt5. In *Proceedings of the first International Workshop on Nakba Narratives as Language Resources*, pages 75–82.

Oanh Tran, Anh Cong Phung, and Bach Xuan Ngo. 2022. Using convolution neural network with bert for stance detection in vietnamese. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7220–7225.