StanceNakba Shared Task: Actor and Topic-Aware Stance Detection in Public Discourse

Kholoud K. Aldous[1], **Md Rafiul Biswas**[2], **Mabrouka Bessghaier**[1], **Shimaa Ibrahim**[1], **Kais Attia**, **Wajdi Zaghouani**[1]
[1]Northwestern University in Qatar, [2]Hamad Bin Khalifa University
mbiswas@hbku.edu.qa,
{kholoud.aldous, mabrouka.bessghaier, shimaa.ibrahim, wajdi.zaghouani}@northwestern.edu
Kais.attia.w@gmail.com

Abstract

We present StanceNakba 2026, a shared task on stance detection in polarized social media discourse related to the Palestinian-Israeli conflict, organized as part of Nakba-NLP 2026 at LREC-COLING 2026. The task introduces two subtasks: Subtask A (Actor-Level Stance Detection), which classifies English social media posts as Pro-Palestine, Pro-Israel, or Neutral; and Subtask B (Cross-Topic Stance Detection), which identifies Favor, Against, or Neither stances in Arabic posts toward two conflict-related topics, normalization with Israel and refugee presence in Jordan. The task is grounded in an annotated dataset of 2,606 social media posts. A total of 7 teams participated in Subtask A and 6 teams in Subtask B. Participating systems primarily fine-tuned Arabic and multilingual transformer-based models, including MARBERT, AraBERT, and DeBERTa-v3 variants, with several teams employing cross-validation, ensemble methods, and topic-conditioned architectures. The best-performing systems achieved a Macro F1 of 0.9620 on Subtask A and 0.8724 on Subtask B, demonstrating that transformer-based approaches are highly effective for conflict-domain stance detection while highlighting persistent challenges in cross-topic generalization and neutral class prediction.

Keywords: Stance Detection, Arabic NLP, Palestinian-Israeli Conflict, Social Media Analysis, X

1. Introduction

Social media platforms now serve as key spaces for political discussion, especially when it comes to deeply polarized issues like the Palestinian–Israeli conflict. The volume, diversity, and emotional intensity of online discussions on this topic present significant challenges for natural language processing (NLP) systems that aim to understand public opinion and political orientation. Stance detection is the task of automatically identifying whether an author is in favor of, against, or neutral toward a given target (Mohammad et al., 2016). It offers a principled framework for analyzing such discourse. However, existing resources and shared tasks have largely focused on English and on politically neutral topics, leaving conflict-related, multilingual, and polarized discourse substantially underexplored.

To address this gap, we present the **StanceNakba 2026 Shared Task**, organized as part of Nakba-NLP 2026: The 2nd International Workshop on Nakba Narratives as Language Resources (Jarrar et al., 2026), co-located with LREC-COLING 2026. StanceNakba 2026 introduces two subtasks to stance detection that distinguish between actor-level political alignments and cross-topic stance patterns in social media posts related to the Palestinian-Israeli conflict and associated regional issues. The task is built on an annotated dataset of 2,606 social media posts in English and Arabic.

The shared task contains two subtasks. **Subtask A** (Actor-Level Stance Detection) is an English-language task that requires systems to identify whether the author of a social media post expresses a *Pro-Palestine*, *Pro-Israel*, or *Neutral* orientation in their general position toward the Palestinian-Israeli conflict. **Subtask B** (Cross-Topic Stance Detection) is an Arabic-language task that requires systems to detect *Favor*, *Against*, or *Neither* stances toward two specific conflict-related topics: normalization with Israel and refugee presence in Jordan. These subtasks enable the investigation of fundamental questions at the intersection of NLP and political discourse analysis: How do general political alignments relate to positions on specific policy issues? Can models learn generalizable stance representations that transfer across different topics within the same conflict domain?

The task attracted broad participation from the research community. A total of 7 teams submitted systems for Subtask A and 6 teams for Subtask B, with two teams participating in both. Participating systems primarily relied on fine-tuned Arabic and multilingual transformer-based models, including MARBERT (Abdul-Mageed et al., 2021), AraBERT (Antoun et al., 2020), and DeBERTa-v3 (He et al., 2021) variants, with the best system achieving a Macro F1 of 0.9620 on Subtask A and 0.8724 on Subtask B. Other techniques such as cross-validation, ensemble methods, Natural Language Inference (NLI) reformulation, and topic-conditioned architec-

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 80–90

tures were explored, demonstrating the richness of approaches that this task inspired.

The remainder of this paper is organized as follows. Section 2 reviews related work on stance detection and Arabic NLP. Section 3 presents the results and system overview for Subtask A, followed by Section 4 for Subtask B. Section 5 concludes with a discussion of findings and directions for future work. Section 6 discusses the limitations of the current task and dataset. Finally, Section 7 provides details about the ethical considerations and dataset availability.

2. Related Work

Stance detection is the task of automatically determining whether the author of a text is in favor of, against, or neutral toward a given target (Mohammad et al., 2016). SemEval-2016 Task 6 (Mohammad et al., 2016) marked a turning point by introducing a widely used benchmark for stance detection in English tweets, covering targets such as political figures, social movements, and ideological positions. Since then, stance detection has been studied across a broad range of social and political domains, with surveys by AlDayel and Magdy (2021) documenting the rapid growth of the field and the shift from traditional feature-based methods toward transformer-based architectures. Transformer-based models have become the dominant paradigm in stance detection. Pre-trained language models such as BERT (Devlin et al., 2019) have been extensively fine-tuned for stance classification, consistently outperforming earlier approaches that relied on lexical features, n-grams, and handcrafted representations. For the Arabic language, dedicated pre-trained models have proven particularly important due to Arabic's morphological richness, dialectal diversity, and diglossia. AraBERT (Antoun et al., 2020) established strong baselines for Modern Standard Arabic NLP tasks, while ARBERT and MARBERT (Abdul-Mageed et al., 2021) extended coverage to dialectal and social media Arabic by pre-training on large-scale Twitter corpora. Arabic stance detection systems now widely rely on these models as their core architecture, as seen in the approaches used by participants in the StanceNakba 2026 shared task.

Arabic stance detection has received growing attention in recent years, though resources remain limited compared to English. The Mawqif dataset (Alturayeif et al., 2022) introduced the first Arabic target-specific stance corpus, comprising over 4,000 tweets annotated for stance, sentiment, and sarcasm across multiple controversial topics. Complementing this, Charfi et al. (2024) introduced a cross-domain, multi-dialectal Arabic stance corpus covering four Arab regions and multiple dialect groups, with over 4,500 annotated sentences balanced across MSA and regional dialects; their AraBERT-based system achieved strong performance and notably outperformed Mawqif-trained models on the neutral stance class, highlighting the importance of class balance in Arabic stance resources. Building on these resources, StanceEval 2024 (Alturayeif et al., 2024) organized the first shared task dedicated to Arabic stance detection, hosted at ArabicNLP 2024. Participating systems mainly fine-tuned Arabic BERT variants, with ensemble methods and multi-task learning emerging as effective strategies for handling class imbalance and dialectal variation. These efforts demonstrated both the feasibility of Arabic stance detection at scale and the challenges that remain, particularly for underrepresented and neutral stance classes.

The StanceNakba 2026 shared task builds upon these foundations while introducing unique challenges tied to the highly polarized nature of discourse surrounding the Palestinian-Israeli conflict. Prior NLP work on conflict-related discourse has examined stance and opinion on related topics, including studies of Twitter polarization around the Israel-Palestine conflict (Imtiaz et al., 2022) and cross-conflict stance correlations (Tao et al., 2024). StanceNakba 2026 is organized as part of the Nakba-NLP 2026 workshop, the second edition of an initiative dedicated to applying NLP tools to the documentation and understanding of Nakba narratives. The shared task distinguishes itself from prior Arabic stance work through its dual-framework design: Subtask A addresses actor-level political alignment in English, classifying authors as Pro-Palestine, Pro-Israel, or Neutral, while Subtask B targets cross-topic Arabic stance detection toward specific conflict-related issues, namely normalization with Israel and refugee presence in Jordan.

Recent work has increasingly emphasized the importance of modeling stance and narrative framing in politically sensitive and conflict-driven contexts. In the Arabic NLP domain, several efforts have focused on capturing polarization, ideology, and media narratives through dedicated datasets and shared tasks. For instance, the FIGNEWS shared task (Zaghouani et al., 2024a) introduced a benchmark for analyzing news media narratives, highlighting the role of framing and bias in shaping public discourse.

Complementary work has explored conflict-related and politically charged discourse on social media platforms. (Shestakov and Zaghouani, 2024) presented a dataset analyzing the digital framing of the Sheikh Jarrah evictions, providing insights into how conflict narratives are constructed and propagated online. Similarly, (Al Heraki and Zaghouani, 2025) examined polarization and misogynistic discourse in Arabic Twitter conversations,

further demonstrating the challenges of modeling stance in emotionally charged environments.

Beyond stance-specific datasets, related efforts in hate speech and propaganda detection have contributed to understanding polarized language and ideological positioning. Resources such as multi-label hate speech corpora (Zaghouani et al., 2024b) and shared tasks on propaganda and subjectivity detection (Hasanain et al., 2024) provide complementary perspectives on how stance, bias, and persuasion interact in online discourse.

These works collectively underscore the need for domain-specific, conflict-aware resources and evaluation frameworks. The StanceNakba shared task builds on this line of research by introducing actor-level and cross-topic stance detection in the context of the Palestinian-Israeli conflict, extending prior work toward more fine-grained and context-sensitive modeling of political stance.

3. Subtask A: Actor-Level Stance Detection

Subtask A frames conflict stance detection at the *actor level*: given a social media post authored by a single user, the model must infer the author's overarching political orientation toward the Palestinian-Israeli conflict. Rather than analyzing individual argumentative moves or claim-specific positions, the task requires models to aggregate signals across the full text and assign a single author-level stance label. This formulation is motivated by the observation that political orientation in polarized discourse is often expressed indirectly through framing choices (which actors are foregrounded, which actions are lexicalized as aggression versus defense, which populations are centered) rather than through explicit opinion markers alone.

3.1. Dataset

The dataset consists of 1,401 English-language posts collected from X (formerly Twitter), all related to the Palestinian-Israeli conflict. Posts were retrieved via the Meltwater media intelligence platform using targeted keyword queries: *"I stand with Palestine"* or *"I stand with Gaza"* for the Pro-Palestine class, *"I stand with Israel"* for the Pro-Israel class, and *"Israel AND Palestine OR Gaza"* for the Neutral class. Labels were assigned based on the query that retrieved each post, yielding a balanced dataset of 467 samples per class (33.3% each).

To ensure data quality, a multi-step preprocessing pipeline was applied. First, stance-signaling seed keywords (e.g., "I stand with Palestine") were stripped to prevent label leakage, along with Twitter mentions, URLs, hashtags, emojis, and special characters, retaining only alphanumeric text and basic punctuation. Posts were then filtered for meaningful content, requiring a minimum of 30 characters and at least five dictionary words of three or more letters, thereby excluding near-empty or uninformative entries. Finally, case-insensitive exact deduplication was performed by comparing lowercase-normalized versions of the cleaned texts, retaining only the first occurrence of each unique post.

Table 1: Subtask A dataset statistics.

Label	Train	Dev	Test	Total
Pro-Palestine	327	70	70	467
Pro-Israel	327	70	70	467
Neutral	326	70	71	467
Total	**980**	**210**	**211**	**1,401**

Data splits: The dataset is partitioned into training (980 samples, 70%), development (210 samples, 15%), and test (211 samples, 15%) splits, each preserving the balanced class distribution. Table 1 summarizes the dataset statistics.

Dataset Examples: The following examples illustrate the range of expression captured by each label:

- **Pro-Palestine:** *"The systematic displacement of Palestinian families from their ancestral homes represents a clear violation of international law and the right of return."*
- **Pro-Israel:** *"Israel's defensive measures are necessary responses to existential threats, ensuring the safety of its citizens against terrorism."*
- **Neutral:** *"The conflict involves competing territorial claims, with both populations having deep historical connections to the region."*

3.2. Setup and Evaluation

The task was organized into two phases:

- **Development phase**: we released the training and development subsets, and participants submitted runs on the development set through a competition on CodaBench.[1]
- **Test phase**: we released the official test subset, and participants were given a few days to submit their final predictions through the same CodaBench competition. Only the latest submission from each team was considered official and was used for the final team ranking.

[1] https://www.codabench.org/

Measures: Subtask A is framed as a three-class classification problem over English social media posts, where participants are asked to build a single unified model that maps an input post to one of three actor-level stance labels: PRO-PALESTINE, PRO-ISRAEL, or NEUTRAL. We measure the performance of the participating systems using **Macro F1**, which computes the unweighted mean of per-class F1 scores, weighting each class equally and thus penalizing poor performance on any individual label, including the minority-leaning NEUTRAL class. Systems are evaluated on a held-out test set, and participants were not permitted to use the test set for model selection or hyperparameter tuning. In addition to Macro F1, Accuracy, Precision, and Recall are also reported on the leaderboard for reference.

3.3. Results and Overview of the Systems

A total of 12 teams submitted runs during the evaluation phase of Subtask A (Actor-Level Stance Detection), of which 7 teams submitted system descriptions and 6 teams submitted system papers. In Table 2, we provide an overview of the participating systems for which a description was submitted. In Table 3, we report the official results for those teams, ranked by Macro F1.

As shown in Table 2, fine-tuning pre-trained transformer-based models is the dominant approach among participating teams, with BERT-based architectures being the most common backbone. Several teams additionally employed cross-validation, ensemble methods, and hyperparameter optimization.

Team **Shroukgbr** (Gabr and Ragab, 2026) achieved the best performance with a Macro F1 of 0.9620, ranking first on the leaderboard. They fine-tuned a single BERT-based transformer model initialized from a publicly available pretrained checkpoint. Preprocessing steps included removing URLs and user mentions, normalizing repeated characters, and truncating inputs to 256 tokens. To address class imbalance, they applied weighted cross-entropy loss. The final configuration used stratified 5-fold cross-validation, an AdamW optimizer with a learning rate of 2e-5, batch size of 16, and 4 training epochs with early stopping.

Team **Yafa** (Zayet et al., 2026) ranked second with a Macro F1 of 0.9525. They fine-tuned MARBERT and ARBERT using a cross-lingual approach. A customized cross-entropy loss with label smoothing and increased attention dropout (0.2) were applied to mitigate overconfidence and overfitting. Preprocessing included lowercase normalization, URL removal, user mention normalization, character repetition removal, whitespace normalization, hashtag normalization, and connected word splitting. The best results were obtained with a learning rate of 2×10^{-5}, weight decay of 0.1, batch size of 4, maximum sequence length of 128, and 8 epochs with early stopping (patience = 2) and an AdamW optimizer.

Team **KUET** (Al Shafi et al., 2026) ranked third with a Macro F1 of 0.9426. They fine-tuned a BERT-based Mixture-of-Experts (MoE) stance classification model built on top of `bert-base-uncased`. The architecture incorporated cue-word masking and contrast-marker amplification within specialized expert modules, along with CNN-based feature extraction with kernel sizes $\{2, 3, 4, 5\}$. They employed 10-fold stratified cross-validation with a weighted ensemble at inference time, where each fold model's contribution was proportional to its validation Macro F1. Label smoothing ($\epsilon = 0.25$) was applied to reduce overconfidence.

Team **KvochurHegel** (Le, 2026) ranked fourth with a Macro F1 of 0.9384. They formulated stance detection as a Natural Language Inference (NLI) task, initializing a Cross-Encoder from the `DeBERTa-v3-base-mnli-fever-anli` checkpoint. Each input text was evaluated against three class-specific engineered hypotheses designed to encode ideological markers. Label smoothing (factor = 0.2) and R-Drop consistency regularization ($\alpha = 4.0$) were applied to mitigate label noise. The model used AdamW with a learning rate of 2e-5, batch size of 16, and early stopping with patience of 2 epochs.

Team **Viva_Palestine** (El-Kassas et al., 2026) ranked fifth with a Macro F1 of 0.9190. They pretrainined `bert-base-uncased` and fine-tuned it by merging training and validation data for the test phase, training for 20 epochs with early stopping (stopping at epoch 9), a maximum token length of 256, and a learning rate of 10^{-5}.

Team **HeatWave**[2] ranked sixth with a Macro F1 of 0.8804. They experimented with multiple transformer-based models and large language models (LLMs), including `microsoft/deberta-v3-large`, `DeBERTa-v3-large-mnli-fever-anli-ling-wanli`, and the election-domain stance model `bert-election2020-twitter-stance-biden-KE-MLM`, the latter achieving the best results upon fine-tuning. They also explored zero-shot prompting with Qwen2.5 and DeepSeek-R1 models, though these yielded lower F1 scores (0.2–0.6). N-fold training with soft ensemble inference was used for the final submission.

Team **The Blackwell Collective** (Shujon et al., 2026) ranked seventh with a Macro F1 of 0.7451. They fine-tuned `microsoft/mdeberta-v3-base` using a Statement Tuning paradigm: inputs were reformatted as cloze templates and

[2]This team didn't submit their system paper

Table 2: Subtask A: Overview of the participating systems. DL: Deep Learning, ML: Classic Machine Learning.

Team	Transformer						DL	Misc.				
	BERT-base	XLM-RoBERTa	ARBERT/MARBERT	DeBERTa-v3	mDeBERTa-v3	KE-MLM BERT	CNN/MoE	Data Augmentation	Preprocessing	Hyperparameter Tuning	Cross-validation	Ensemble Methods
Shroukgbr (Gabr and Ragab, 2026)	✓	✓	✓					✓	✓	✓	✓	✓
Yafa (Zayet et al., 2026)			✓							✓		
KUET (Al Shafi et al., 2026)	✓						✓			✓	✓	✓
KvochurHegel (Le, 2026)				✓						✓		
Viva_Palestine (El-Kassas et al., 2026)	✓									✓		
HeatWave				✓		✓				✓	✓	✓
The Blackwell Collective (Shujon et al., 2026)					✓					✓		

Table 3: Official results for Subtask A, ranked by Macro F1 score. Leaderboard ranks are shown alongside overall submission ranks (in parentheses). Of the 12 total test submissions, only 7 teams provided a model description and are included here.

Team	Rank	Macro F1
Shroukgbr	1 (#1)	0.9620
Yafa	2 (#2)	0.9525
KUET	3 (#3)	0.9426
KvochurHegel	4 (#4)	0.9384
Viva_Palestine	5 (#6)	0.9190
HeatWave	6 (#7)	0.8804
The Blackwell Collective	7 (#11)	0.7451

the `[MASK]` token's hidden state was contrasted against three learnable class prototypes via a softmax head. Three auxiliary modules were incorporated: Prototype Contrastive Learning (PCL) with $K = 3$ global prototypes at temperature $\tau = 0.1$, Topic-Conditioned Layer Normalization (T-CLN) conditioning feature distributions on topic identity, and R-Drop consistency regularization via symmetric KL divergence. Training followed a two-stage schedule: backbone frozen for 2 epochs, then full fine-tuning for 8 epochs, using AdamW with a learning rate of 10^{-5} for the backbone and 10^{-4} for the heads.

4. Subtask B: Cross-Topic Stance Detection

Subtask B focuses on cross-topic stance detection, where systems are required to determine the stance expressed in a text toward a given target topic. In the context of the StanceNakba Shared Task, this subtask centers on politically sensitive discourse related to the Palestinian cause and regional political dynamics.

4.1. Dataset

We used a subset of the MARASTA dataset (Charfi et al., 2024) for this subtask. MARASTA is a cross-domain, multidialectal Arabic stance corpus designed to support stance detection across multiple dialectal regions of the Arab world. The dataset covers four major dialectal regions, namely Maghreb, Egypt, the Levant, and the Gulf, representing the primary dialect groups in Arabic.

Dataset Overview The full MARASTA corpus contains more than 4,500 sentences annotated with stance toward eight controversial topics distributed across the four Arab regions. Each sentence is labeled according to its stance toward a given topic using three categories: pro (favor), against, or neutral. The dataset was constructed with careful balancing strategies. Each topic contains approximately 500-700 sentences, with a relatively even distribution across the three stance classes. In addition, the corpus maintains a balance between MSA and dialectal Arabic, with roughly half of the sentences written in MSA and the remaining half in the dialect associated with the topic's region.

Data Collection The data were collected from online platforms such as Twitter and YouTube, where discussions on controversial socio-political topics are common. The collected content could be either

current or historically relevant within the past 15 years. For each region (Maghreb, Egypt, the Levant, and the Gulf), the two most highly discussed topics were selected. Candidate sentences were retrieved using seed keywords associated with each topic. To do so, the authors used Python scripts leveraging the platforms' APIs to retrieve tweets and YouTube comments containing the selected keywords. The collected sentences were then manually filtered to ensure they were written in Arabic (MSA or dialect), grammatically correct, relevant to the topic, and expressed a stance (pro, against) or a neutral position, allowing the dataset to capture naturally occurring stance expressions across different dialects and communication styles.

Data Annotation The annotation process followed a multi-stage quality control procedure. Each sentence was independently annotated by two annotators to determine its stance with respect to the associated topic. In cases of disagreement, the instance was reviewed by a third annotator who resolved the conflict and determined the final label. The resulting annotations demonstrate substantial inter-annotator agreement, with an overall Cohen's Kappa score of 0.84 and regional agreement scores ranging from 0.80 to 0.89, indicating high annotation reliability.

Topic Selection for the StanceNakba Shared Task Since the StanceNakba Shared Task focuses on political discourse and narratives related to the Palestinian cause and regional political dynamics, we selected the two topics from the MARASTA corpus that are most relevant to this theme: "Normalization with Israel" and "Refugee/Immigrant Presence in Jordan". The dialectal content associated with these topics reflects two distinct Arabic regional varieties: the "Normalization with Israel" topic contains Gulf Arabic and MSA, while the "Refugee/Immigrant Presence in Jordan" topic contains Levantine Arabic (Jordanian/Palestinian) and MSA. These topics capture key socio-political debates in the Arab world that are closely connected to discussions surrounding Palestine, regional diplomacy, and displacement in the Middle East. The main characteristics of the subset used in this shared task are summarized below:

- **Source:** Arabic social media posts (X) discussing Palestinian-Israeli conflict-related issues.
- **Total Size:** 1,205 annotated samples across two topics
- **Topics Covered:**

Topic 1: التطبيع مع إسرائيل (Normalization with Israel)

- Total: 577 samples
  - Against: 198 (34.3%)
  - Neutral: 208 (36.0%)
  - Pro: 171 (29.6%)

Topic 2: وجود اللاجئين والمهاجرين في الأردن من فلسطين، سوريا، العراق (Refugee/Immigrant Presence in Jordan)

- Total: 628 samples
  - Against: 228 (36.3%)
  - Neutral: 163 (26.0%)
  - Pro: 237 (37.7%)

Data Split

- Training: 843 samples (70%)
- Development: 181 samples (15%)
- Test: 181 samples (15%)

Dataset Examples
Topic: Normalization with Israel

- **Pro:** "الجامعة العربية قالت إنها لا ترى أن #التطبيع مع #إسرائيل خطوة ضد القضية الفلسطينية"
  English: The Arab League said it does not see normalization with Israel as a step against the Palestinian cause.
- **Against:** "إن قدرة الدولة على التطبيع جهارًا نهارًا مع النظام الإسرائيلي تتماشى مع قوة واستقرار نظامها المستبد"
  English: A state's ability to normalize openly with the Israeli regime aligns with the strength and stability of its authoritarian system.
- **Neutral:** "كشف تقرير نشرته البوابة الإسرائيلية لشؤون الزراعة أنّه منذ سنة ونصف تجري في أوروبا اجتماعات بين ممثلين إسرائيليين واماراتيين"
  English: A report published by the Israeli agricultural portal revealed that meetings between Israeli and Emirati representatives have been taking place in Europe for a year and a half.

Topic: Refugee/Immigrant Presence in Jordan

- **Pro:** " لا مكان للعنصرية بالأردن اي شخص داخل الأردن يعامل معاملة ابن البلد وهاذا الشخص يمثل كل أردني شريف"
  English: There is no place for racism in Jordan. Any person inside Jordan is treated like a son of the country, and this person represents every honorable Jordanian.

4.2. Setup and Evaluation

Subtask B follows the same two-phase setup (development and test) and evaluation protocol as Subtask A (see Section 3.2), with Macro F1 as the primary metric computed over the three labels: FAVOR, AGAINST, and NEITHER.

4.3. Results and Overview of the Systems

A total of 8 teams submitted runs during the evaluation phase of Subtask B (Cross-Topic Stance Detection), of which 6 teams submitted system descriptions and system papers. In Table 4, we provide an overview of the participating systems for which a description was submitted. In Table 5, we report the official results for those teams, ranked by Macro F1.

As shown in Table 4, fine-tuning Arabic pretrained transformer models is the dominant strategy, with several teams incorporating cross-validation, ensemble methods, and topic-conditioned input formulations to improve cross-topic generalization.

Team **Viva_Palestine** (El-Kassas et al., 2026) achieved the best performance in Subtask B with a Macro F1 of 0.8724, ranking first on the leaderboard. They fine-tuned `UBC-NLP/MARBERT` by merging the training and validation sets to form an expanded training set of 1,024 records. The input was formatted as a topic–sentence pair truncated to 128 tokens. The best hyperparameters were: learning rate of 2×10^{-5}, batch size of 8, 4 training epochs, warmup ratio of 0.1, weight decay of 0.01, and the AdamW optimizer.

Team **EGCSS** (Qindeel et al., 2026) ranked second with a Macro F1 of 0.8607. They fine-tuned `bert-base-arabertv02-twitter` with a classification head, incorporating topic descriptions concatenated to the input to provide cross-topic signal. They experimented with data augmentation via back-translation, Claude-generated tweets, and the ArabicStanceX dataset, but none of these yielded improvements over the base configuration and were excluded from the final submission. The best hyperparameters were: learning rate of 2×10^{-5}, maximum length of 64, 10 epochs with a linear learning rate schedule, and early stopping based on the labeled validation set.

Team **U4RASD** (Hamdan et al., 2026) ranked third with a Macro F1 of 0.8601. Their final system fine-tuned MARBERTv2 with dialect-aware LLM-based data augmentation, using Gemini 3 Flash Preview to generate three paraphrased variants of each training sample while preserving both stance label and dialectal register. They also investigated contrastive learning, multi-task learning with LLM-generated auxiliary labels, counterfactual augmentation, and zero-shot prompting, none of which improved over the dialect-aware augmentation baseline. Final model selection was based on the best Macro F1 on the development set using stratified 5-fold cross-validation.

Team **The Resistant Word** (Bahgat et al., 2026) ranked fourth with a Macro F1 of 0.8562. They employed a multi-model ensemble with 5-fold stratified cross-validation, fine-tuning four pre-trained Arabic transformer models: MARBERT, AraBERT Large, XLM-RoBERTa Base, and CAMeLBERT-Mix. Each model received a topic-sentence pair truncated to 128 tokens. Prior to training, random oversampling was applied to balance the training set. All models were trained for 5 epochs with a learning rate of 2×10^{-5}, batch size of 16, and BF16 mixed precision. A custom weighted cross-entropy loss was applied, with class weights computed inversely proportional to class frequency. Final predictions were produced by averaging softmax probabilities across all 20 runs (4 models $\times$ 5 folds).

Team **A2NLP** (Nairat and Nairat, 2026) ranked fifth with a Macro F1 of 0.8483. They fine-tuned `aubmindlab/bert-base-arabertv02-twitter` with a prompt-based input formulation that explicitly conditions stance prediction on the target topic. Preprocessing was tailored to Arabic social media and included emoji normalization, URL and mention removal, diacritic removal, Alef variant unification, and whitespace normalization. Stratified 5-fold cross-validation was employed with model selection based on validation Macro F1. A weighted cross-entropy loss addressed class imbalance. The best hyperparameters were: learning rate of 2×10^{-5}, batch size of 16, up to 10 epochs with early stopping (patience = 2).

Team **The Blackwell Collective** (Shujon et al., 2026) ranked sixth with a Macro F1 of 0.7407. They applied the same Statement Tuning architecture as in Subtask A, fine-tuning `microsoft/mdeberta-v3-base` with cloze-style input templates and a `[MASK]`-based prototype classifier. Topic-Conditioned Layer Normalization (T-CLN) served as the primary mechanism for cross-topic generalization, dynamically conditioning feature distributions on a learnable 64-dimensional topic embedding. Prototype Contrastive Learning (PCL) and R-Drop regularization were also applied. Training followed the same two-stage schedule as in Subtask A: backbone frozen for 2 epochs, then full fine-tuning for 8 additional epochs using AdamW.

5. Conclusion and Future Work

We presented an overview of the StanceNakba 2026 Shared Task, which addresses stance detection in polarized social media discourse on the Palestinian-Israeli conflict. The task introduced a dual-framework approach consisting of two sub-

Table 4: Subtask B: Overview of the participating systems. DL: Deep Learning, ML: Classic Machine Learning.

Team	Transformer					Misc.					
	AraBERT (Twitter)	MARBERT	XLM-RoBERTa	mDeBERTa-v3	CAMeL-BERT	Data Augmentation	Preprocessing	Hyperparameter Tuning	Cross-validation	Ensemble Methods	Topic Conditioning
Viva_Palestine (El-Kassas et al., 2026)	✓	✓					✓	✓			✓
EGCSS (Qindeel et al., 2026)	✓					✓	✓	✓	✓		✓
U4RASD (Hamdan et al., 2026)		✓				✓			✓		
The Resistant Word (Bahgat et al., 2026)		✓	✓		✓			✓	✓	✓	
A2NLP (Nairat and Nairat, 2026)	✓					✓	✓	✓	✓		✓
The Blackwell Collective (Shujon et al., 2026)				✓			✓	✓			✓

Table 5: Official results for Subtask B, ranked by Macro F1 score. Leaderboard ranks are shown alongside overall submission ranks (in parentheses). Of the 8 total test submissions, only 6 teams provided a model description and are included here.

Team	Rank	Macro F1
Viva_Palestine	1 (#1)	0.8724
EGCSS	2 (#2)	0.8607
U4RASD	3 (#3)	0.8601
The Resistant Word	4 (#4)	0.8562
A2NLP	5 (#5)	0.8483
The Blackwell Collective	6 (#7)	0.7407

tasks: **Subtask A** (Actor-Level Stance Detection), identifying whether authors express Pro-Palestine, Pro-Israel, or Neutral orientations; and **Subtask B** (Cross-Topic Stance Detection), detecting Favor, Against, or Neither stances toward specific conflict-related topics, namely normalization with Israel and refugee presence in Jordan.

The task attracted considerable interest from the research community. A total of 7 teams made official submissions for Subtask A, and 6 teams for Subtask B, with two teams participating in both subtasks. For both subtasks, the majority of systems fine-tuned pre-trained Arabic transformer models, with BERT-based architectures — particularly MARBERT, AraBERT, and DeBERTa-v3 variants — being the most common backbone. Top-performing systems on Subtask A achieved Macro F1 scores as high as 0.9620, while Subtask B results were competitive with the best system reaching a Macro F1 of 0.8724, reflecting the added difficulty of cross-topic generalization. Several teams incorporated other techniques such as cross-validation, ensemble methods, topic-conditioned input formulations, and consistency regularization to improve performance.

Future editions of the StanceNakba shared task could explore more fine-grained stance categories and extend coverage to additional conflict-related topics and Arabic dialects. Increasing dataset size, particularly for underrepresented stance classes, and developing evaluation protocols that better capture cross-topic generalization will be important directions. Future work could also investigate the role of large language models (LLMs) under zero-shot and few-shot settings.

6. Limitations

Several participating teams noted difficulty predicting the neutral or "neither" class, likely due to its inherent ambiguity. Additionally, the cross-topic nature of Subtask B introduces domain shift between training and test topics, which current transformer-based models do not fully address. Expanding the dataset with more balanced label coverage and including a broader range of topics and dialectal varieties would help mitigate these limitations in future iterations of the task.

7. Ethical Considerations & Dataset Availability

7.1. Data Source and Consent

The datasets used in the StanceNakba Shared Task consist of publicly available posts collected from Twitter X in accordance with the platform's

terms of service. Only publicly accessible content was included, and no attempt was made to access private accounts, restricted material, or deleted posts.

Usernames, profile information, and other direct identifiers were removed. The dataset does not include user metadata beyond the textual content necessary for stance classification. In compliance with platform policy, only post IDs and annotation labels will be released where required.

7.2. Sensitive and Political Content

This shared task focuses on discourse related to the Palestinian Israeli conflict and related regional issues. This topic is politically sensitive and may involve references to violence, displacement, discrimination, or traumatic events.

The dataset may contain polarized language and emotionally charged expressions. The purpose of the shared task is scientific analysis of stance detection and cross topic generalization. It is not intended to endorse, legitimize, or amplify any political position. Results should not be interpreted as normative judgments.

7.3. Risk of Harm and Misuse

Automated stance detection systems applied to political discourse may be misused for surveillance, political profiling, targeted persuasion, or repression. We explicitly discourage the use of models trained on this dataset for surveillance or discriminatory purposes.

Because Subtask A involves inference of general political alignment, predictions should not be treated as reliable indicators of an individual's beliefs or identity. Model outputs are probabilistic classifications based on limited textual evidence and may be incorrect.

7.4. Annotation Bias and Subjectivity

Stance annotation is inherently interpretive. Although detailed guidelines were developed for the Pro Palestine, Pro Israel, Neutral, Favor, Against, and Neither labels, annotators may bring implicit cultural, political, or linguistic biases.

To mitigate this risk, annotators were trained using standardized definitions, disagreements were adjudicated, and balanced label distributions were maintained where feasible. Nevertheless, residual bias may remain in both the English and Arabic datasets. Models trained on this data may learn annotation artifacts rather than purely linguistic signals of stance.

7.5. Representation and Sampling Limitations

The datasets consist of a limited number of annotated English and Arabic social media posts. They are not representative of all political perspectives, demographic groups, or geographic populations. Social media users are themselves not representative of broader societies.

Accordingly, findings should not be generalized to entire populations or interpreted as reflecting majority opinion. Claims about cross topic generalization should be framed cautiously.

7.6. Cross Lingual and Cross Cultural Considerations

Subtask A and Subtask B involve different languages and sociopolitical contexts. Linguistic markers of stance may vary across dialects, regions, and political cultures. Models may encode unintended cultural assumptions.

Participants are encouraged to report per class performance, conduct qualitative error analysis, and critically examine model behavior across topics and languages.

7.7. Annotator Well Being

Exposure to polarized or conflict related discourse may cause emotional fatigue. Annotation processes should include reasonable workload distribution and allow annotators to opt out of reviewing content they find distressing.

7.8. Data Availability

The dataset is released for non commercial research purposes. Users must comply with platform terms of service, avoid attempts at deanonymization, and refrain from surveillance or discriminatory applications. The dataset can be accessed upon request through the following form: https://forms.gle/YUFdA16R6HkSZjp88

Acknowledgment

This shared task was made possible by the National Priorities Research Program grant NPRP14C-0916-210015 from the Qatar National Research Fund (QNRF), part of the Qatar Research, Development and Innovation Council (QRDI).

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: long papers), pages 7088–7105.

Huda Al Heraki and Wajdi Zaghouani. 2025. Analyzing digital polarization on hijab: A dataset of annotated youtube comments. In *Proceedings of ICWSM 2025*.

Abdullah Al Shafi, Md. Milon Islam, Sk. Imran Hossain, and K. M. Azharul Hasan. 2026. KUET at StanceNakba shared task: StanceMoE: Mixture-of-experts framework for stance detection. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

Abeer AlDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Nora Alturayeif, Hamzah Luqman, Zaid Alyafeai, and Asma Yamani. 2024. Stanceeval 2024: The first arabic stance detection shared task. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 774–782.

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.

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.

Mohammed Bahgat, Doaa Salah, and Sarah Yassine. 2026. The Resistant Word at StanceNakba shared task: A topic-aware model for cross-topic stance detection. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

Anis Charfi, Mabrouka Bessghaier, Andria Atalla, Raghda Akasheh, Sara Al-Emadi, and Wajdi Zaghouani. 2024. Stance detection in arabic with a multi-dialectal cross-domain stance corpus. *Social Network Analysis and Mining*, 14.

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 (long and short papers)*, pages 4171–4186.

Wafaa S. El-Kassas, Enas A. Hakim Khalil, and Enas M.F. El Houby. 2026. Viva_Palestine at 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.

Shrouk Anwar Gabr and Mohamed Ibrahim Ragab. 2026. Shroukgbr at StanceNakba shared task: Transformer-based ensemble learning for actor-level stance detection in Palestinian–Israeli social media discourse. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

Nancy Hamdan, Aya Jouni, Aya Saïd, and Fadi A. Zaraket. 2026. U4RASD at StanceNakba shared task: Data augmentation and auxiliary objectives for arabic stance detection. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

Maram Hasanain, Reem Suwaileh, Sander Weering, Chengkai Li, Tommaso Caselli, Wajdi Zaghouani, Alberto Barrón-Cedeño, Preslav Nakov, and Firoj Alam. 2024. Overview of the clef-2024 checkthat! lab task on subjectivity in news articles. In *CLEF 2024 Working Notes*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

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

Mustafa Jarrar, Mo El-Haj, Amal Haddad, Serin Atiani, Shadi Abudalfa, Khalil Sima'an, Paul Rayson, and Camille Mansour, editors. 2026. *Proceedings of the second International Workshop on Nakba Narratives as Language Resources*. Association for Computational Linguistics, Spain.

Minh-Hoang Le. 2026. KvochurHegel at StanceNakba shared task: Robust stance detection with regularized natural language inference. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

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.

Alaa Nairat and Aysar Mahmoud Nairat. 2026. A2NLP at StanceNakba shared task: Fine-tuned AraBERT for topic-based arabic stance detection. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

Asmaa Qindeel, Toka Khaled, Batool Balah, Eman Elrefai, and Mahmoud Fawzi. 2026. EGCSS at StanceNakba shared task: Cross-topic arabic stance detection for two middle east issues. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

Alexey Shestakov and Wajdi Zaghouani. 2024. Analyzing conflict through data: A dataset on the digital framing of sheikh jarrah evictions. In *Proceedings of the Workshop on NLP for Political Sciences at LREC-COLING 2024*.

Md. Shakhoyat Rahman Shujon, MD Jahid Hasan Jim, Md. Milon Islam, Md Rezwanul Haque, and Fakhri Karray. 2026. The Blackwell Collective at StanceNakba shared task: PAST-TIDE: Prototype-anchored statement tuning with topic-invariant normalization for stance detection. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

Yiyao Tao, Hengyu Zhang, Babli Dey, Selenge Tulga, Hanjia Lyu, and Jiebo Luo. 2024. In the eyes of the bystander: Are the stances on different conflicts correlated? In *2024 IEEE International Conference on Big Data (BigData)*, pages 7193–7201. IEEE.

Wajdi Zaghouani, Mustafa Jarrar, Nizar Habash, Houda Bouamor, Imed Zitouni, Mona Diab, Samhaa El-Beltagy, and Mohammed AbuOdeh. 2024a. The fignews shared task on news media narratives. In *Proceedings of LREC-COLING 2024*.

Wajdi Zaghouani, Hamdy Mubarak, and Md Rafiul Biswas. 2024b. So hateful! building a multi-label hate speech annotated arabic dataset. In *Proceedings of LREC-COLING 2024*.

Tasnim Zayet, Osama Hamed, and Tasneem Duridi. 2026. Yafa at StanceNakba shared task: Actor-level stance detection using cross-lingual approach. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.